TECHNISCHE
UNIVERSITÄT
DARMSTADT

UNTERSTÜTZUNG DES RESSOURCEN-BASIERTEN LERNENS IN ONLINE COMMUNITIES – AUTOMATISCHE ERSTELLUNG VON GROSSTAXONOMIEN IN VERSCHIEDENEN SPRACHEN

Vom Fachbereich Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

DIPL.-INFORM. RENATO DOMÍNGUEZ GARCÍA

Geboren am 17. September 1982 in San José, Costa Rica

Referent: Prof. Dr.-Ing. Ralf Steinmetz
Korreferent: Prof. Dr.-Ing Ulrik Schroeder

Tag der Einreichung: 07. Dezember 2012
Tag der Disputation: 04. Februar 2013

Hochschulkennziffer D17
Darmstadt 2013

KURZFASSUNG

Die sich stetig verändernden beruflichen Rahmenbedingungen und die immer kürzer werdende Gültigkeit einmal erworbenen Wissens verlangen flexible Formen des Wissens- und Kompetenzerwerbs. Das in Bildungseinrichtungen angeeignete Wissen reicht nicht mehr ein Leben lang. Vielmehr besteht insbesondere im Arbeitsprozess zunehmend die Notwendigkeit, sich abhängig von der konkreten Problemstellung situativ Wissen anzueignen. Man spricht von selbstgesteuertem Lernen, da Lernende für ihre Lern- bzw. Wissenserwerbsprozesse selbst verantwortlich sind. Gleichzeitig hat sich das World Wide Web zu einer der wichtigsten Quellen beim Wissenserwerb entwickelt. Das selbstgesteuerte Lernen mit Hilfe von Ressourcen aus dem Internet wird auch Ressourcen-basiertes Lernen bezeichnet.

Eine der größten Herausforderung im Ressourcen-basierten Lernen ist es, relevante Web-Ressourcen im Web zu finden. Suchmaschinen werden sehr häufig verwendet, liefern aber praktisch keine Hilfestellung bei der Auswahl und Beurteilung gefundener Ressourcen. Empfehlungssysteme (engl. Recommender Systems) können grundsätzlich hilfreich sein, um für die jeweilige Situation und den jeweiligen Lernenden relevanten Ressourcen zu finden. Lernende können davon profitieren, dass sie auf Wissensressourcen hingewiesen werden, die andere Lernende, die einen ähnlichen Wissensbedarf besitzen, verwendet haben. In größeren Gruppen oder in einer Community sind für die eigene Lernaufgabe relevante Ressourcen mit hoher Wahrscheinlichkeit bereits von anderen Personen gefunden worden.

Ziel dieser Arbeit war es, das Ressourcen-basierte Lernen innerhalb einer Community von Lernenden zu unterstützen, indem Lernende situationsbezogen auf Wissensressourcen hingewiesen werden, die andere Community-Mitglieder bereits verwendet haben.

Zur Erreichung dieses Ziels wurde das Anwendungsszenario am Beispiel der CROKODIL-Plattform, eine Plattform zur Unterstützung des Ressourcen-basierten Lernens, untersucht. Die Untersuchung ergab, dass Benutzer oftmals nicht auf interessante Ressourcen hingewiesen werden können, wenn sie unterschiedliche Terminologien bei der Verschlagwortung von beim Lernen genutzten Ressourcen verwenden. Basierend auf dieser Feststellung wurde ein Konzept entwickelt, welches die Lücken in den von den Benutzern verwendeten Terminologien mittels der Verwendung einer Taxonomie schließt. Die Analyse ergab weiterhin, dass das Anwendungsszenario dadurch gekennzeichnet ist, dass die Benutzer aktuelle Begriffe in mehreren Sprachen als Schlagworte verwenden. Taxonomien, die diese Schlagworte in Beziehung zueinander setzen wollen, müssen daher dadurch charakterisiert sein, dass sie sehr aktuell sind und im mehreren Sprachen vorliegen. Diese Anforderungen können von existierenden manuell erzeugten Taxonomien nicht erfüllt werden.

Daher wurden in der Arbeit mit TaxWikiHeur.KOM und TaxWikiML.KOM zwei Verfahren konzipiert und implementiert, die weitestgehend sprachunabhängig aus der Online Enzyklopädie Wikipedia Taxonomien generieren, indem sie Kategorienpaare aus der Wikipedia in Hyponymie- und Nicht-Hyponymiebeziehungen klassifizieren. Diese Verfahren zeichnen sich dadurch aus, dass sie keine externen, manuell erzeug-

ten Wissensbasen verwenden. Damit besteht keine Notwendigkeit einer manuellen Pflege von Taxonomien für neue Wissensbereiche. Das Verfahren TaxWikiML.KOM erweitert das Verfahren TaxWikiHeur.KOM und behebt einige der bei der Evaluation von TaxWikiHeur.KOM erkannten Mängel. Die Evaluation der Verfahren hat insgesamt gezeigt, dass trotz des Verzichtes auf eine externe Wissensbasis die Güte der Taxonomien sehr gut ist. Die Verwendung der Verfahren erfolgte in fünf Sprachen, so dass der Nachweis der sprachunabhängigen Nutzbarkeit ebenfalls erfolgte.

Das Verfahren TaxWikiML.KOM wurde in der Arbeit weiterhin verwendet, um innerhalb der CROKODIL-Lernumgebung automatisch Beziehungen zwischen von den Benutzern verwendeten Schlagworten zur Beschreibung der im Lernprozess genutzten Ressourcen zu ergänzen. Es konnte zum einen anhand dreier Korpora aus dem Anwendungsfeld der Ressourcen-basierten Lernens nachgewiesen werden, dass die Dichte des semantischen Netzes, die zur Speicherung der Daten (Ressourcen, Tags und Benutzer) benutzt wird, durch das implementierte Konzept größer wird, womit Empfehlungssysteme umfangreichere Informationen zur Generierung von Empfehlungen zur Verfügung stehen, die auch solche Ressourcen anderer Lernender empfehlen können, die mit einer unterschiedlichen Terminologie beschrieben sind. Der positive Einfluss von mittels TaxWikiML.KOM ergänzten Hyponymiebeziehungen zwischen Schlagworten auf die Güte von Empfehlungssystemen wurde in einer weiteren Evaluation anhand des State-of-the-Art Verfahrens FolkRank zusätzlich nachgewiesen.

Schließlich wurde das FReSET-Tool zur Evaluation von Empfehlungssystemen entwickelt. Das Tool wurde bereits in verschiedenen Arbeiten zur Evaluation verwendet, da es einen standardisierten Vergleich von Empfehlungssystemen ermöglicht.

ABSTRACT

Due to constantly changing professional environments and the decrease in the half-life of acquired knowledge, flexible forms of knowledge and skills acquisition are required. Nowadays, the knowledge acquired in educational institutions no longer last a lifetime. Rather, there is an increasing need (especially in work processes) for the personal acquisition of knowledge depending on specific tasks. This is called self-directed learning, as learners are responsible for their learning processes. At the same time, the World Wide Web has become one of the most important sources for knowledge acquisition. Self-directed learning using resources from the Internet is also called resource-based learning.

One of the biggest challenges in resource-based learning is finding relevant web resources on the Web. Search engines are very often used for this purpose, but they do not provide assistance in the selection of found resources. Recommender systems can be helpful to find relevant resources. Learners can benefit from resources that other learners with similar knowledge requirements have already found. In larger groups or in a community, there is a high probability that relevant resources have already been found by other people.

The goal of this thesis is to support resource-based learning within a community of learners by recommending knowledge resources that other community members have already found.

To achieve this objective, the application scenario and an example implementation, CROKODIL, were investigated. The investigation revealed that the recommendation of interesting resources is often impossible, if the users use different terminologies for the tagging of resources. Based on this observation, a concept was developed that fills the gaps in the terminology used by the users through the use of a taxonomy. The analysis also reveals that the application scenario is characterized by current terms in multiple languages which are used as tags. A taxonomy used for the purpose of finding relationships between tags must, therefore, be characterized by the fact that it is up-to-date and available in multiple languages. These characteristics make manually created taxonomies unsuitable.

Therefore, two approaches, TaxWikiHeur.KOM and TaxWikiML.KOM, were designed and implemented in order to generate large-scale taxonomies from the online encyclopedia Wikipedia in multiple languages. This is done by classifying pairs of categories from the Wikipedia in hyponymy and non-hyponymy relationships. These methods are characterized by the fact that they do not use external, manually created knowledge bases. Thus there is no need for the manual maintenance of taxonomies for new knowledge fields. TaxWikiML.KOM extends TaxWikiHeur.KOM and fixes some of the recognized shortcomings in the evaluation of TaxWikiHeur.KOM. The evaluation of the whole process has shown that even if no external knowledge base was used, the quality of the taxonomies was still very good. The approaches were evaluated in five different languages, in order to show the language-independency of the approaches.

TaxWikiML.KOM was also used within CROKODIL to complement automatically generated relations between tags used by the users to describe the resources in their learning processes. Based on three corpora obtained from the application scenario, the evaluation could show that the density of the network grew using the implemented concept. Therefore, recommender systems have more information available to generate recommendations and this can be used for recommendations to learners using different terminologies. Additionally, the positive effect on the quality of recommender systems due to hyponymy relations between tags found by TaxWikiML.KOM was demonstrated in a further evaluation based on a state-of-the-Art algorithm.

Finally, the FReSET tool for the evaluation of recommender systems was developed. FReSET can be used for the evaluation of recommender systems as it allows a standardized and thus comparable evaluation of recommender systems.

INHALTSVERZEICHNIS

1

EINFÜHRUNG

»Ein Anfang ist kein Meisterstück, doch guter Anfang halbes Glück.«

— Anastasius Grün

1.1 MOTIVATION

Die sich stetig verändernden beruflichen Rahmenbedingungen und die immer geringer werdende Gültigkeit einmal erworbenen Wissens [39, 56, 164] verlangen flexible Formen des Wissens- und Kompetenzerwerbs [148]. Das in Bildungseinrichtungen angeeignete Wissen reicht nicht mehr ein Leben lang. Vielmehr besteht insbesondere im Arbeitsprozess zunehmend die Notwendigkeit, sich abhängig von der konkreten Aufgabenstellung situativ Wissen anzueignen. Die Lernenden sind dann für ihre Lern- bzw. Wissenserwerbsprozesse selbst verantwortlich und können entscheiden, wann sie was, wo und wie lernen möchten. Man spricht von *selbstgesteuertem Lernen*. Gleichzeitig hat sich das World Wide Web (WWW) zu einer der wichtigsten Quellen beim Wissenserwerb entwickelt. Das WWW enthält verschiedenste Quellen, wie Onlineenzyklopädien, Weblogs oder Nachrichtenportale, aber auch frei verfügbare Lernressourcen (Open Educational Ressources) und wissenschaftliche Publikationen. Teilweise sind diese Ressourcen zu Lernzwecken didaktisch aufbereitet, überwiegend aber nicht. Das WWW zeichnet sich zudem dadurch aus, dass Informationen zu aktuellen Themen vorliegen, die in Lehrbüchern noch nicht aufgenommen sind.

Das selbstgesteuerte Lernen mit Hilfe von solchen Ressourcen, wie sie im Internet zu finden sind, wird auch Ressourcen-basiertes Lernen (RBL) genannt und wurde von Norbert Meder in [97] als „ein Sich-verfügbar-Machen von Informationen und Wissensbeständen bei aktuellen Problemen“ beschrieben. Selbstgesteuertes Ressourcen-basiertes Lernen ist mit einer Vielzahl von Herausforderungen für den Lernenden verbunden [145]. Daher wurden Anwendungen entwickelt, um den Lernenden individuell im Ressourcen-basierten Lernen zu unterstützen [19].

Eine der größten Herausforderung im Ressourcen-basierten Lernen ist es, relevante Web-Ressourcen im Web zu finden [41]. Suchmaschinen werden sehr häufig verwendet, liefern aber praktisch keine Hilfestellung bei der Auswahl und Beurteilung gefundener Ressourcen. Empfehlungssysteme [131] (engl. Recommender Systems) können grundsätzlich hilfreich sein, um für die jeweilige Situation und den jeweiligen Lernenden relevanten Ressourcen zu finden [128]. Allerdings muss beachtet werden, dass an Empfehlungssysteme im E-Learning besondere Anforderungen bestehen. Während Systeme für Produktempfehlungen, wie beispielsweise in Amazon[1], die Empfehlung ähnlicher Produkte zum Ziel haben, so ist das im E-Learning nicht immer sinnvoll. Es gibt zum Beispiel Lernende mit verschiedenen Kenntnisstufen. Anfänger brauchen Lernressourcen, die einen groben Überblick über ein bestimmtes

1 https://www.amazon.de/ - Zugriff am 14.11.2012

Thema geben, während Experten Lernressourcen brauchen, die vertieft in das Thema eingehen [39].

1.2 ZIEL, ANSATZ UND BEITRÄGE DER ARBEIT

In vielen Lernszenarien können sich Lernende gegenseitig unterstützen. Das gilt auch für das Ressourcen-basiertes Lernen. Lernende können davon profitieren, dass sie auf Wissensressourcen hingewiesen werden, die andere Lernende, die einen ähnlichen Wissensbedarf besitzen, verwendet haben. In größeren Gruppen oder in einer Online Community (wie z.B. Social Bookmarking Applikationen wie delicious[2] oder GroupMe[3]) sind für die eigene Lernaufgabe relevante Ressourcen mit hoher Wahrscheinlichkeit bereits von anderen Personen gefunden worden. Zielsetzung dieser Arbeit ist es, gerade im Ressourcen-basierten Lernen dem Lernenden die Ressourcen, die innerhalb einer Community bereits verwendet wurden, situationsabhängig zugänglich zu machen. Die Arbeit betrachtet zusammenfassend als generelles Szenario das kollaborative Ressourcen-basierte Lernen in Online Communities.

Die Arbeit analysiert dieses Szenario am Beispiel der CROKODIL-Lernumgebung [8], die Ressourcen-basiertes Lernen unterstützt, und zeigt bestehende Schwächen der zur Verfügung stehenden Anwendung auf. So können Lernende heute, obwohl Empfehlungssysteme in der CROKODIL-Lernumgebung realisiert sind, nur teilweise von den Ressourcen anderer Community-Mitglieder profitieren. Dies resultiert insbesondere aus der Tatsache, dass die Lernenden verschiedene Terminologien bei der Verschlagwortung von Ressourcen verwenden. Anfänger kennen die spezifische Terminologie des Themas, für das sie sich interessieren nicht. Experten dagegen kennen und benutzen die spezifische Terminologie des Gebietes, um sich präzise auszudrücken.

Der Ansatz dieser Arbeit besteht darin, eine Taxonomie dazu zu verwenden, um diese Unterschiede in der von den Benutzern verwendeten Terminologie zu überbrücken. Mittels der Taxonomie sollen Beziehungen zwischen den von den Lernenden zur Verschlagwortung von Ressourcen verwendeten Begriffen ergänzt werden. Damit stehen zusätzliche Informationen zur Verfügung, die verwendet werden sollen, um verbesserte Empfehlungssysteme zu realisieren und damit dem Lernenden die Ressourcen anderer Community-Mitglieder zugänglich zu machen.

Das Szenario des Ressourcen-basierten Lernens in Communities stellt an die Taxonomie die Anforderung, dass sie einerseits aktuelle Begriffe aus nahezu beliebigen Wissensdomänen enthalten muss, um Lernenden in ihren akuten Lernaufgaben zu unterstützen, und andererseits in mehreren Sprachen vorliegen muss, da Lernende häufig Ressourcen in unterschiedlichen Sprachen verwenden und ebenso Schlagworte aus unterschiedlichen Sprachen zur Auszeichnung der Ressourcen nutzen. Zur Erstellung solcher aktueller, umfassender Taxonomien in mehreren Sprachen verfolgt diese Arbeit den Ansatz die Wikipedia, als umfassende mehrsprachige Onlineenzyklopädie, als Wissensbasis zu verwenden. Die vorliegende Arbeit umfasst folgende Beiträge:

- Das Szenario des kollaborativen Ressourcen-basierten Lernens in Online-Communities wird anhand des Beispiels der CROKODIL-Lernumgebung analysiert und es werden bestehende Herausforderungen identifiziert.

2 http://delicious.com/ - Zugriff am 14.11.2012

3 http://groupme.org/GroupMe/ - Zugriff am 14.11.2012

- Basierend auf dieser Analyse wird ein Konzept zur Bereitstellung von zusätzlichen Relationen zwischen den von Lernenden zur Verschlagwortung verwendeten Begriffen auf Basis von Taxonomien entwickelt. Das Ziel ist hierbei Lernenden die Ressourcen anderer Community-Mitglieder zugänglich zu machen. Die Anforderungen an die genutzte Taxonomie werden ebenfalls bestimmt.
- Es werden zwei Verfahren zur sprachunabhängigen Generierung von Taxonomien auf Basis der Wikipedia entworfen, die die zuvor bestimmten Anforderungen erfüllen, konzipiert, implementiert und evaluiert.
- Das Konzept zur Bereitstellung von zusätzlichen Relationen wird als Erweiterung der CROKODIL-Lernumgebung implementiert.
- Die Verwendung der Taxonomien zur Ergänzung von Relationen zwischen Schlagworten in der CROKODIL-Lernumgebung wird evaluiert.
- Ein Framework zur Evaluation von auf Folksonomien basierenden Empfehlungssystemen wird konzipiert, implementiert und verwendet, um den Nutzen der Verwendung von Relationen zwischen Schlagworten anhand eines Standardempfehlungssystems zu evaluieren.

1.3 GLIEDERUNG DER ARBEIT

Die vorliegende Arbeit gliedert sich wie folgt: Nach dieser Einleitung erfolgt in Kapitel 2 die Beschreibung des Anwendungsszenarios Ressourcen-basiertes Lernen und der für das Verständnis der Arbeit notwendigen Grundlagen. Die wichtigsten Begriffe werden definiert. Kapitel 3 fasst verwandte Arbeiten zu dem in dieser Arbeit behandelten Themen und Ansätzen zusammen. In Kapitel 4 wird das Anwendungsszenario des kollaborativen Ressourcen-basierten Lernens detailliert anhand der CROKODIL-Lernumgebungen analysiert und es werden die Herausforderungen bestimmt. Die Zielsetzung der Arbeit und das Konzept zur Bestimmung von Relationen zwischen Schlagworten auf Basis von Taxonomien und deren Verwendung in Empfehlungssystemen werden vorgestellt. Kapitel 5 stellt zwei sprachunabhängige Methoden vor, mit deren Hilfe sich taxonomische Beziehungen aus der Wikipedia bestimmen lassen, und evaluiert sie. Kapitel 6 beschreibt die Umsetzung des zuvor vorgestellten Konzeptes in der CROKODIL-Plattform. Anschließend wird in Kapitel 7 der Nutzen der ergänzten Relationen evaluiert. Kapitel 8 fasst den Inhalt dieser Arbeit zusammen und schließt die vorliegende Arbeit mit einem Ausblick auf zukünftige Forschungsarbeiten ab.

2

GRUNDLAGEN

»Man muss sicher auf festem Boden gehen können, ehe man mit dem Seiltanzen beginnt.«

— Henri Matisse

Dieses Kapitel führt die Terminologie ein, die im Rahmen dieser Arbeit benutzt wird. Im ersten Abschnitt werden grundlegende Arbeiten und Begriffe zum Thema Information Retrieval und kollaboratives Ressourcen-basiertes Lernen vorgestellt. Da sich diese Arbeit mit Taxonomien zur Unterstützung des Ressourcen-basierten Lernens beschäftigt, soll im dritten Abschnitt ein Überblick über verschiedene Möglichkeiten der Wissensrepräsentation gegeben werden. Abschließend wird auf Wikipedia als Wissensquelle eingegangen, weil diese im Rahmen dieser Arbeit Wikipedia als Wissensquelle zur Generierung einer Taxonomie benutzt wird.

2.1 RESSOURCEN-BASIERTES LERNEN UND LERNRESSOURCEN

Das Internet hat sich zu einer wichtigen Quelle von im Lernprozess zu verwendenden Ressourcen entwickelt. Heutzutage ist ein großer Teil des menschlichen Wissens digital über das Internet verfügbar. Beispielsweise werden Bücher über Initiativen wie Google Books[1] digitalisiert. Darüber hinaus können Bibliotheken ihre Sammlungen in die Google Buchsuche aufnehmen lassen. Ein anderes Beispiel für Institutionen, die Ressourcen im Internet zur Verfügung stellen, sind wissenschaftliche Gesellschaften und Verlage wie Association for Computing Machinery (ACM)[2] bzw. Springer Link[3], die schon lange digitale Kopien von wissenschaftlichen Beiträgen anbieten. Aber nicht nur Bücher und wissenschaftliche Publikationen sind online verfügbar, sondern beispielsweise auch Anleitungen für Haushaltsgeräte oder Computer-Treiber. Zusätzlich sind in den letzten Jahren die sogenannten *Web 2.0 Technologien*, wie Weblogs (Blogs), Soziale Netzwerke oder Foren entstanden. Diese Technologien erlauben Internet-Nutzern, eigenes Wissen im Web verfügbar zu machen und mit Interessierten zu interagieren. Diese Entwicklung ist in informellen Lernsettings, insbesondere in Lernsettings, in denen Lernende die Organisation und Planung ihres Lernprozesses selbst übernehmen, relevant.

Meder definiert in [97] die Art des Lernens mit Ressourcen z.B. aus dem Internet als „ein Sich-verfügbar-Machen von Informationen und Wissensbeständen bei aktuellen Problemen“. Diese Art des Lernens hat Rakes [123] als einen Lernmodus, bei dem Lernende durch eigene Interaktion mit einer großen Vielfalt an Ressourcen anstatt durch Frontalunterricht lernen, bezeichnet. Allerdings schlägt Rakes vor, dass Lehrende eine Vorauswahl an Ressourcen vornehmen sollten und nicht die Studenten das gesamte Netz durchstöbern sollen. Tergan [157] bezog die Definition von Rakes vor

1 http://books.google.com - Zugriff am 14.11.2012
2 http://dl.acm.org - Zugriff am 14.11.2012
3 http://www.springerlink.com - Zugriff am 14.11.2012

allem auf Hypertextumgebung und Internetressourcen und erwartete von Lernenden die selbstständige Suche nach Lernressourcen im Web.

Heutige Schätzungen[4] gehen davon aus, dass das Web mehr als 7,9 Billionen Webseiten umfasst, und diese Anzahl wächst täglich. Auch wenn nur ein geringer Teil dieser Webseiten relevante und wertvolle Informationen, die für das Lernen verwendet werden können, enthalten, ist es immer noch eine unüberschaubare Anzahl von Ressourcen. Damit ergeben sich für das Lernen mit diesen Ressourcen Herausforderungen: Viele Ressourcen sind nicht für das Lernen aufbereitet. Relevante Informationen sind zum Beispiel oft über verschiedene Ressourcen verteilt. Die Unerfahrenheit von Lernenden ist ein weiteres Problem, denn sie können zumindest am Anfang einer Recherche nicht über die Vertrauenswürdigkeit und Relevanz von Ressourcen urteilen. Weitere Probleme werden von Tergan in [157] und Naumann in [107] angesprochen. Tergan spricht von struktureller und konzeptueller Desorientierung, die auftreten, wenn ein Lernender seine Recherche nicht strukturieren kann (z.B. wenn er nicht weiß, wie er am besten die Suche startet oder wenn er Schwierigkeiten hat beim Auffinden bereits besuchter Webseiten) bzw. wenn er neue Informationen nicht aufnehmen kann (z.B. weil Vorwissen fehlt). Naumann wiederum spricht von der kognitiven Mehrbelastung, wenn Lernende ihre Konzentration nicht nur für das Lernen, sondern für andere Aufgaben wie das Finden, Bewerten oder Speichern von Ressourcen verwenden müssen.

Ressourcen-basiertes Lernen mit Ressourcen aus dem Internet findet in sehr vielen Szenarien statt: Wenn Studenten einen Vortrag im Rahmen einer Gruppenarbeit erstellen möchten und Informationen im Netz suchen, wenn ein Schüler ein Biologie-Referat vorbereiten soll, wenn Mitarbeiter einer Firma eine Geschäftsreise nach Indien planen und sich über Land und die Kultur informieren wollen oder wenn Wissenschaftler an einem Beitrag arbeiten und nach verwandten Arbeiten recherchieren. An diesen Beispielen werden zwei Aspekte deutlich: erstens, dass das Vorwissen der Personen nicht ausreicht, um einen aktuellen Informationsbedarf zu decken, sodass sie selbstständig nach digitalen Ressourcen suchen müssen und zweitens, dass Lernende zwar die Aufgabe alleine bearbeiten, aber oft einer größeren Gruppe angehören. Beispielsweise gehört ein Schüler zu einer Klasse, ein Studenten besucht eine Vorlesung zusammen mit anderen Studenten, ein Mitarbeiter ist einer von vielen in einer Firma und ein Forscher einer von vielen in einer Forschergruppe. Aus diesem Grund stellte Tergan fest, dass Lehrende Lernenden Werkzeugen zur Unterstützung des Ressourcenbasierten Lernens anbieten sollten. In [19] diskutiert Böhnstedt weitere Definitionen des Ressourcen-basierten Lernens im Web und definiert Ressourcen-basiertes Lernen als „Form des Lernens, bei welcher der aktuelle Informationsbedarf durch die selbständige Interaktion mit einer Vielzahl verschiedener digitaler Lernressourcen gedeckt wird“. Diese Definition beschreibt sehr genau das Ressourcen-basierte Lernen, wie es im Rahmen dieser Arbeit benutzt wird, betrachtet allerdings nicht die oben erwähnte zweite Tatsache: Da Lernende in vielen Szenarien Mitglieder einer größeren Gruppe sind, gibt es eine hohe Wahrscheinlichkeit, dass andere Mitglieder relevante oder ähnliche Ressourcen bereits gefunden haben. Daher wird in dieser Arbeit Ressourcen-basiertes Lernen wie folgt definiert:

4 http://www.worldwidewebsize.com/ - Zugriff am 14.11.2012

Definition 1 (Ressourcen-basiertes Lernen (RBL) in Online-Communities) *RBL in Online Communities wird definiert als Form des Lernens, bei welcher Lernende ihren aktuellen Informationsbedarf durch selbständige Interaktion mit einer Vielzahl verschiedener digitaler Lernressourcen decken. Lernende gehören dabei einer Community an, deren andere Mitglieder ebenfalls durch selbständige Interaktion mit Lernressourcen lernen. Die gesammelten Lernressourcen stehen allen Mitgliedern der Community zur Verfügung.*

Diese Definition schließt weder die Anleitung durch einen Experten oder Lehrer noch die Kollaboration zwischen den Mitgliedern der Community aus, fokussiert aber auf das selbstgesteuerte Suchen und Lernen eines Lernenden.

In dieser Definition kommt das Konzept „Lernressource“ vor. Bevor die Herausforderungen im Ressourcen-basierten Lernen in Communities analysiert werden, soll dieser Begriff an dieser Stelle nochmals erklärt werden. In der Literatur sind die Begriffe „Lernressourcen“ und „Lernobjekte“ eng miteinander verknüpft, oft werden sie sogar synonym benutzt. Die Tatsache, dass sich sehr viele Forscher mit Lernressourcen beschäftigt haben, hat dazu geführt, dass sich keine klare Definition des Begriffs „Lernobjekt“ gebildet hat. Die existierenden Definitionen hängen in der Regel vom Anwendungsszenario ab. Scholl hat in [145] verschiedene Definitionen diskutiert und folgerte, dass Forscher in der Vergangenheit folgende Merkmale in den Vordergrund der Definition von Lernobjekten stellen:

- **Granularität**, also die Feinkörnigkeit der Lernobjekte, je nachdem, ob ein Lernobjekt aus vielen anderen kleinen besteht. Hier haben sich Autoren wie Wiley [165], Polsani [116], Boyle [22] oder Meyer [100] mit der Frage beschäftigt, ob Fragmente von Lernobjekten auch Lernobjekte sind und wie sich aus diesen Fragmenten neue Lernobjekte erstellen lassen.

- **Wiederverwendbarkeit**, die lange Zeit im Fokus der Forschung war, definiert, inwieweit sich Lernobjekte für verschiedene Zwecke adaptieren lassen. Autoren wie Polsani [116], Hörmann [59], Meyer [100] und Zimmermann [174] legten einen großen Fokus auf die Wiederverwendbarkeit von Lernobjekten.

- **Lernabsicht**, d.h. inwieweit die Absicht mit dem ein Objekt erstellt wurde, seine Eigenschaften als Lernobjekt betrifft. Beispielsweise definierte L'Allier [80] verschiedene Kriterien, die Ressourcen erfüllen müssen, um als Lernobjekte zu gelten. Eines dieser Kriterien war u.a. die Tatsache, dass ein Lernziel definiert ist. Darüber hinaus muss dieses Ziel durch Deckung eines Informationsbedarfs erreichbar sein und es muss bewertbar sein, ob und ab wann das Ziel erreicht wurde. Polsani [116] und Littlejohn [84] definierten Lernobjekte allerdings weniger strikt, da eine zu große Fixierung auf Lernziele die Wiederverwendbarkeit beeinträchtigen könnte.

- **das Beinhalten von Metadaten**, um die Suche, Katalogisierung und den Nutzen von Lernobjekten zu unterstützen. Hodgins klassifizierte in [58] Lernobjekte je nach der Art der Metadaten, mit der sie versehen sind.

- **ihr Inhaltsmodell**, je nachdem, wie der Inhalt des Lernobjekts eingebunden ist. Bekannte Inhaltsmodelle sind das Cisco-Inhaltsmodell [14] und das Inhaltsmodell von Hodgins [58].

- **im Lebenszyklus**, unterteilt in Erstellung, Wiederverwendung, Benutzung und Bereitstellung. Die Unterstützung des Lebenszyklus von Lernobjekten haben sich Autoren wie Downnes [38], Rensing et al. [126] und Lehmann [82] vorgenommen.

Allerdings stellt Polsani in [116] fest, dass diese vielen Definitionen von Lernobjekten nicht konsistent sind und sich zum Teil widersprechen. Darüber hinaus existieren Formate wie das Learning Object Metadata (LOM)[5] (LOM) oder das Shareable Content Object Reference Model (SCORM)[6] (SCROM), um Lernobjekte zu beschreiben oder auszutauschen. Im Ressourcen-basierten Lernen, wie es in dieser Arbeit betrachtet wird, spielen Web-Ressourcen eine große Rolle, insofern schränken die Definitionen von Lernobjekten aus der Literatur das Ressourcen-basierte Lernen zu sehr ein. Aus diesem Grund werden im Rahmen dieser Arbeit Lernressourcen wie folgt definiert:

Definition 2 (Lernressourcen) *Lernressourcen sind digitale (Web-)Ressourcen, die vom Lernenden im Ressourcen-basierten Lernen genutzt werden können.*

Lernressourcen können z.B. Webseiten, Videos, Bilder oder Blog-Einträge sein. Darüber hinaus deckt diese Definition die Definitionen aus vorherigen Arbeiten ab, wenn die Lernobjekte digital vorliegen und über das Web verfügbar sind. Laut Definition 2 müssen Lernressourcen weder explizit für das Lernen aufbereitet noch durch Metadaten beschrieben sein oder andere Eigenschaften besitzen.

2.2 INFORMATION RETRIEVAL UND MASCHINELLES LERNEN

Information Retrieval (IR) (auf Deutsch Informationsrückgewinnung) wird von Manning et al. in [91] wie folgt definiert: „*Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)*“. Danach ist Information Retrieval als das Finden von unstrukturierten Dokumenten in einer großen Menge von Dokumenten, um einen Informationsbedarf zu decken, definiert. Im Folgenden werden verschiedene Konzepte aus diesem Gebiet eingeführt und insbesondere auf die Evaluation von Systemen für Information Retrieval eingegangen.

2.2.1 *Information Retrieval*

2.2.1.1 *Informationsbedarf*

Der Informationsbedarf eines Benutzers bezeichnet den Wunsch nach Informationen zu einem gegebenen Thema, um z.B. eine gegebene Aufgabe zu lösen [91]. Man unterscheidet zwischen Informationsbedarf und Anfrage. Eine Anfrage beschreibt den Versuch eines Benutzers, seinen Informationsbedarf mitzuteilen [91]. Dies kann in textueller Form (z.B. SQL-Anfrage[7]), implizit (z.B. mittels Analyse des Benutzer-

5 `http://www.adlnet.gov/Technologies/scorm/SCORMSDocuments/20044thEdition/` - Zugriff am 14.11.2012

6 `http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf` - Zugriff am 14.11.2012

7 Standard Query Language (SQL)

verhaltens oder seiner gespeicherten Dateien) oder auf andere Weisen geschehen(z.B. mit „Like"-Knöpfen in Facebook[8])

2.2.1.2 *Relevanz*

Eine Ressource oder ein Dokument ist relevant für einen Benutzer, wenn die Ressource wertvolle Informationen bzgl. des Informationsbedarfs eines Benutzers [91] enthält. Allerdings muss man an dieser Stelle anmerken, dass die Relevanz immer eine Portion Subjektivität hat und dass die Qualität des Retrieval von der Anfrage des Benutzers abhängt.

2.2.1.3 *Ranking*

Informell kann ein Ranking als eine Liste von Entitäten gemäß einem Ranking-Kriterium bezeichnet werden. Beispielsweise kann ein Ranking von Ressourcen eine Liste von Ressourcen sein, in dem die Ressourcen in absteigender Reihenfolge gemäß Relevanz den Informationsbedarf eines Benutzers decken.

Formell ist ein Ranking ein Tupel von Entitäten in geordneter Reihenfolge. Entitäten sind eine Menge von Dingen, die gerankt werden und die miteinander vergleichbar sein können auf der Basis eines Ranking-Kriteriums, das durch die binäre Relation $\triangleleft$ ausgedrückt wird.

$$R = \{(e_0,...,e_n) | (e_0,...,e_n) \in P(E) \wedge \forall k = \{0,...,n-1\} \Rightarrow e_{k+1} \triangleleft e_k\}$$

E stellt eine Menge von Entitäten dar, die gerankt werden sollen.
$P(E)$ Menge von Permutationen von E
$\triangleleft$ Totale, reflexive und transitive binäre Relation basierend auf Ranking-Kriterien

2.2.1.4 *Empfehlung*

Eine Empfehlung von Entitäten sind Vorschläge. Üblicherweise werden diese Vorschläge von einem Empfehlungssystem generiert. Entitäten können alle möglichen Ressourcen, Benutzer oder Dinge sein. Beispielsweise beschreibt Koren in [77] ein Film-Empfehlungssystem. In [3] stellen Adomavicious et al. verschiedene traditionelle Empfehlungssysteme dar. Das Ziel von traditionellen Empfehlungssystemen ist die Bestimmung einer Ratingfunktion R mit folgender Signatur:

$$R : Benutzer \times Entität \rightarrow Rating$$

Die Ratingfunktion ist partiell, da kein Benutzer alle Entitäten gespeichert hat. Ein Empfehlungssystem muss die Ratingfunktion total machen, also unbekannte Ratings raten. In Folksonomie-Anwendungen 2.3.8 ist $Rating \in \{0,1\}$, je nachdem, ob ein Benutzer eine Ressource getaggt hat oder nicht. Die Abschätzung von unbekannten Ratings stellt ein Ranking dar. Aus diesem Grund lassen sich Empfehlungssysteme als Rankingsysteme ansehen. Ein Empfehlungssystem prognostiziert Entitäten basierend auf Informationen über sie und auf dem Profil des Benutzers [64]. In Abschnitt 3.1 wird auf die verschiedenen Empfehlungssysteme näher eingegangen.

8 `http://www.facebook.com` - Zugriff am 14.11.2012

2.2.2 *Maschinelles Lernen*

Der Einsatz von Methoden des maschinellen Lernens für die Klassifizierung von Relationen zwischen den Konzepten stellt einen Schwerpunkt dieser Arbeit dar. Im folgenden Kapitel sollen die Prinzipien erklärt werden, wie solche Verfahren evaluiert werden. Mitchell hat in [103] maschinelles Lernen wie folgt definiert: „A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with Experiences E“. Es geht also darum, dass ein Computer-Programm von Erfahrungen E lernt, wie eine Aufgabe (oder Menge von Aufgaben) T gelöst wird, sodass sie bzgl. einer gegebenen Metrik bzw. Maß besser abschneidet. Beispielsweise kann ein Schachspieler-Computer-Programm seine Gewinnquote (P) beim Schachspielen (T) durch Erfahrungen verbessern, wenn es immer wieder gegen sich selbst spielt (E).

In dieser Arbeit wird maschinelles Lernen im Rahmen von binären Klassifikationsaufgaben angewendet. Eine Klassifikationsaufgabe bezeichnet die Aufgabe der Klassifizierung von Instanzen in einer gegebenen diskreten Menge von möglichen Kategorien [103]. Bei binären Klassifikationsaufgaben geht es darum, zwischen genau zwei Kategorien zu unterscheiden. Die Klassifizierungsaufgabe übernimmt ein Klassifikator. Der Klassifikator entscheidet anhand einer gegebenen Beispielmenge, wie eine neue Instanz klassifiziert wird. Angewendet auf die Definition von maschinellem Lernen stellt man fest, dass die Aufgabe (T) darin besteht, zwischen zwei Kategorien c_1 und c_2 zu unterscheiden, die Erfahrungen (E) kommen aus einer gegebenen Beispielmenge, genannt Trainingskorpus, und die Performanz (P) wird anhand der im nächsten Abschnitt (2.2.3) vorgestellten Metriken gemessen.

Konkret entscheidet der Klassifikator basierend auf sogenannten Features, ob eine Instanz zu c_1 oder c_2 gehört. Ein Feature wird auf eine Instanz angewendet und liefert als Ergebnis jedes Features einen Zahlenwert zurück. Anschließend, wenn alle Features zu einer Instanz berechnet wurden, werden die Werte in einen sogenannten Featurevektor eingetragen. Der Featurevektor wird dabei als Repräsentation der Instanz angesehen und kann benutzt werden, um ähnliche Instanzen oder wiederkehrende Muster zwischen den Instanzen zu berechnen. Die Erstellung eines Featurevektors wird in Abbildung 1 dargestellt.

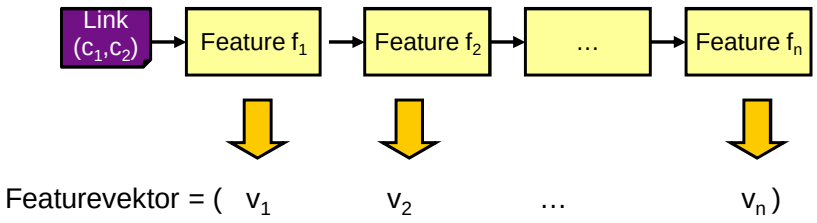

Abbildung 1: Erstellung des Featurevektors

2.2.3 *Evaluationsmaße*

Die im Rahmen dieser Arbeit entwickelten Verfahren werden mit Hilfe von Maßen aus dem Information Retrieval evaluiert: Precision, Recall und F-Maß. Diese Maße helfen dabei, die Güte der Verfahren aus mehreren Blickwinkeln zu beurteilen.

Gegeben sei ein Klassifikator k, der eine (z.B. binäre) Klassifikation c_1 oder c_2 vornimmt. Der Klassifikator klassifiziert Entitäten $e = \{e_1,e_2,...,e_n\}$ in den beiden Klassen c_1 oder c_2 ein. Bei diesem Prozess kann der Klassifikator allgemeine Fehler begehen, d.h. er sortiert eine Entität e_i in die falsche Klasse ein. Abhängig vom Ergebnis der Klassifikation sowie von der tatsächlichen Klasse der Entität können vier Fälle unterschieden werden:

1. Richtig-positiver Fall (engl. true positive, t_p): Eine Entität e_i der Klasse c_1 wird korrekt als c_1 markiert.
2. Richtig-negativer Fall (engl. true negative, t_n): Eine Entität e_i der Klasse c_2 wird korrekt als c_2 markiert.
3. Falsch-positiver Fall (engl. false positive, f_p): Eine Entität e_i der Klasse c_1 wird fälschlicherweise als c_2 markiert.
4. Falsch-negativer Fall (engl. false negative, f_n): Eine Entität e_i der Klasse c_2 wird fälschlicherweise als c_1 markiert.

Oft benutzt man eine sogenannte *Konfusionsmatrix*, um die Ergebnisse der Klassifikation darzustellen. Ein Beispiel wird in Tabelle 1 dargestellt.

Tabelle 1: Beispiel einer Konfusionsmatrix

	Der Link gehört zur Klasse c_1	Der Link gehört zur Klasse c_2
Als c_1 klassifiziert	Richtig-positiv (t_p)	Falsch-positiv (f_p)
Als c_2 klassifiziert	Falsch-negativ (f_n)	Richtig-negativ (t_n)

Aus den Werten der Konfusionsmatrix lassen sich zwei zentrale Kennzahlen zur Evaluation eines Klassifikators errechnen [90]:
Recall (auch Sensitivität oder Trefferquote genannt), die als Anteil der korrekt als c_1 klassifizierten Entitäten an der Gesamtheit der tatsächlich existierenden c_1 Entitäten definiert werden. Recall entspricht der bedingten Wahrscheinlichkeit:

$$P(\text{richtig als } c_1 \text{ erkannt} \mid \text{alle tatsächlichen existierenden } c_1\text{-Entitäten}) = \frac{tp}{tp+fn}$$

Precision (auch Relevanz, positiver Vorhersagewert, Genauigkeit genannt), der Anteil der korrekt als c_1 klassifizierten Entitäten an der Gesamtheit der als c_1 erkannten Entitäten. Precision entspricht der bedingten Wahrscheinlichkeit:

$$P(\text{richtig als } c_1 \text{ erkannt} \mid \text{alle als } c_1\text{-erkannten Entitäten}) = \frac{tp}{tp+fp}$$

Recall und Precision stehen oft in Konflikt zueinander: Ein höherer Recall eines Klassifikators bedeutet, dass der Großteil der tatsächlichen c_1-Entitäten vom Klassifikator als c_1 korrekt klassifiziert wurde. Darunter kann aber die Precision leiden, da der Klassifikator für diesen Zweck viele c_2-Entitäten u.U. klassifizieren müsste. Umgekehrt kann eine hohe Precision zu einem schlechten Recall führen. Ein Klassifikator, der „auf Nummer sicher“ geht und nur Entitäten als c_1 klassifiziert, wenn er sich sehr sicher ist, produziert hohe Precision-Werte. Dies führt aber auch dazu,

dass viele „unsichere“ c_1-Entitäten übersprungen werden, was wiederum zu einer niedrigeren Precision führt.

Je nach Anwendungsszenario können abhängig vom Schwerpunkt der Suche Suchergebnisse mit höherem Recall oder mit höherer Precision bevorzugt werden, indem dem Recall oder der Precision höhere Gewichtung zugeordnet wird [133]. Aus dem Precision und dem Recall kann ein sogenanntes **F-Maß** berechnet werden, das ein kombiniert gewichtetes harmonisches Maß für Recall und Precision darstellt. Im Rahmen dieser Arbeit wird die sogenannte F_1-Maß (Precision und Recall werden jeweils mit dem Wert 1 gewichtet) wie folgt berechnet:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

In allgemeinerer Form lautet die Formel für F_α (mit $\alpha \geqslant 0$) [133]:

$$F_\alpha = \frac{(1+\alpha) * Precision * Recall}{\alpha * Precision + Recall}$$

F_2 gewichtet z.B. den Recall doppelt so stark wie die Precision, bei $F_{0,5}$ hingegen nimmt die Precision eine doppelt höhere Gewichtung im Vergleich zum Recall ein. Van Rijsbergen postuliert in [133], dass F-Maß ein Maß für die Effektivität der Informationsgewinnung aus Sicht eines Benutzers ist.

Ein wichtiger Unterschied zwischen Recall und Precision ist der Fakt, dass das Recall nicht vom Verhältnis zwischen den positiven und negativen Fällen in der Testmenge abhängt. Der Recall beschränkt sich nur auf die positiven Fälle (c_1) und es spielt keine Rolle, ob die positiven Fälle in der Testmenge unter- oder überrepräsentiert sind. Die Berechnung der Precision dagegen bezieht sowohl die positiven als auch die negativen Fälle in die Berechnung mit ein, was dazu führt, dass die Precision durch das Verhältnis zwischen positiven und negativen in der Testmenge beeinflusst wird.

2.2.4 *Evaluationsverfahren*

Für die Evaluation von Verfahren im Information Retrieval werden oft Varianten der sogenannten Kreuzvalidierung (engl. cross-validation) genutzt [103]. Mittlerweile hat sich die sogenannte k-fache stratifizierte Kreuzvalidierung (engl. K-Fold Cross-Validation) als aus statischer Sicht beste Wahl herauskristallisiert [16]. Im Rahmen dieser Arbeit werden die Ergebnisse des Verfahrens des maschinellen Lernens mittels einer zehnfachen stratifizierten Kreuzvalidierung überprüft. Die Vorgehensweise wird in Abbildung 2 gezeigt: Der gesamte Korpus wird in zehn Stichproben zerlegt. Davon werden neun Proben (90 % der Proben, hier in Grün dargestellt) als Trainingsdaten benutzt und die zehnte Probe (10% der Proben, hier in Rot dargestellt) als Testdaten. Der Evaluationsvorgang wird 10 Mal durchlaufen, wobei jede Stichprobe genau einmal als Testprobe eingesetzt wird. Abschließend werden die erhaltenen Ergebnisse über eine Mittelwertbildung bzw. über eine andere Kombinierungsmethode zusammengeführt, um ein einheitliches gesamtes Ergebnis zu erreichen.

Der Vorteil dieser Methode liegt darin, dass alle Stichproben sowohl für das Training als auch für die Validierung des Verfahrens benutzt werden und jede Stichprobe genau einmal als eine Testprobe auftritt. Durch den Einsatz der zehnfachen stratifizierten Kreuzvalidierung wird der Einfluss von zufälligen Ausreißern vermindert sowie eine klare Trennung zwischen Training- und Test-Daten erreicht [76].

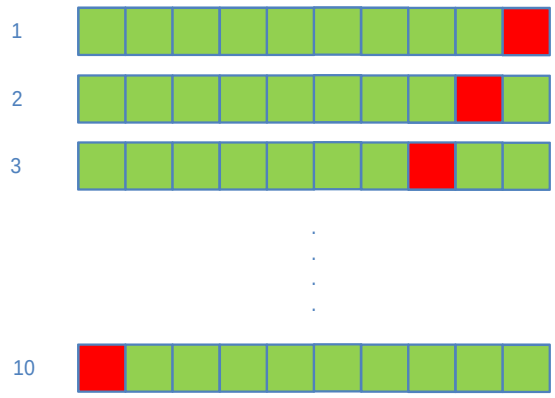

Abbildung 2: 10-fache stratifizierte Kreuzvalidierung

2.3 WISSENSREPRÄSENTATION

In dieser Arbeit werden Fachbegriffe aus dem Gebiet der Wissensrepräsentation und der Computerlinguistik verwendet. In diesem Kapitel sollen daher die zum Verständnis der Arbeit notwendigen Fachbegriffe definiert und erklärt werden.

2.3.1 *Begriffe*

Im Rahmen dieser Arbeit sind *Begriffe* Wörter oder zusammengesetzte Wörter, die eine syntaktische und semantische Einheit bilden. Beispiele von Begriffen sind „Maschine“, „Informationstechnologie“, „Trauer“, „Telefon“ und „Löwe“.

2.3.2 *Konzepte*

Es existieren sehr viele Definitionen eines Konzepts. An dieser Stelle sollen drei Definitionen vorgestellt und diskutiert werden.

Definition von Grabrilovitch and Markovitch [46]: *„Concepts are the basic units of meaning that serve humans to organize and share their knowledge.“*

Konzepte sind also demnach die „Grundeinheiten“ der Bedeutung, die von Menschen benutzt werden, um ihr Wissen zu organisieren und zu teilen. Bei dieser Definition liegt der Fokus in der Bedeutung eines Wortes. Diese Definition erlaubt durch die Benutzung des Begriffes „Grundeinheit des Wissens“ einen großen Raum an Interpretationen. Es ist hier sehr schwer zu sagen, ob ein gegebener Begriff eine „Grundeinheit des Wissens“ darstellt. Das liegt daran, dass das Wissen von Menschen nicht objektiv messbar ist.

Definition aus Wikipedia.org[9]: *„Ein Konzept ist ein Ergebnis des Instruments des Prozesses kognitiver Konzeption von Begriffen mit Sprache, der gleichzeitig eine Aussage zum Begriff enthält.“*

Diese Definition enthält Begriffe wie „kognitive Konzeption“, die aus der Psychologie kommen. Bei dieser Definition wird ein Konzept als ein Ergebnis der Konzeption deklariert, allerdings ist es mit ihrer Hilfe sehr schwierig zu beurteilen, ob ein bestimmter Begriff ein Konzept darstellt.

9 `http://de.wikipedia.org/wiki/Konzeption` - Zugriff am 14.11.2012

Definition aus WordNet[10]: *„A concept is an abstract or general idea inferred or derived from specific instances"*

Diese Definition ist abstrakter als die zwei vorherigen Definitionen. Sie stellt weder erklärende Beschreibungen vor noch geht sie im Detail auf die spezifischen Instanzen ein. Zwar werden hier die Eigenschaften eines Konzepts definiert, sie erlaubt aber z.B. im Rahmen dieser Arbeit Kategorien in Wikipedia als Konzepte zu sehen und die darin enthaltenen Artikel als Instanzen zu sehen.

Zum Schluss bleibt zu erwähnen, dass Konzepte eindeutig sind. Beispielsweise stellt jede Bedeutung des Begriffes „Bank", das Finanzinstitut und die Sitzgelegenheit, ein einzigartiges und unabhängiges Konzept dar.

2.3.3 *Relationen zwischen Konzepten*

Semantische Relationen: Bevor die verschiedenen Modelle zur Wissensrepräsentation vorgestellt werden, sollen hier einige der in solchen Modellen vorkommenden semantischen Beziehungen zwischen Konzepten vorgestellt werden.

Synonymie: Synonymie bezeichnet die inhaltliche Überstimmung zwischen zwei oder mehreren Begriffen. Synonyme Begriffe beschreiben dasselbe Konzept. Beispiele für Synonyme sind die Begriffe „Wagen" und „Auto". Beide beschreiben ein Fahrzeug mit vier Rädern.

Antonymie: Antonymie besteht zwischen zwei Begriffen, wenn einer der beiden Begriffe das Gegenteil des anderen ist. Beispielsweise sind „Kälte" und „Wärme", „weiß" und „schwarz" Antonymien.

Meronymie: Meronymie stellt eine Teil-Ganzes-Relation zwischen Begriffen dar. Meronymie-Relationen gibt es zwischen „Finger" und „Hand" oder '"Tür" und „Haus".

Hypernymie und Hyponymie: Als Hyponymie wird in der Linguistik eine Relation zwischen zwei Begriffen bezeichnet, bei der ein Begriff in einen anderen Begriff eingeschlossen ist. Nach Cruse [33] wird ein Konzept c_1 als Hyponym von c_2 bezeichnet (und umgekehrt wird c_2 als Hyperonym von c_1 bezeichnet), wenn jedes Individuum von c_1 auch ein Individuum von c_2 ist, aber nicht umgekehrt. Beispielsweise ist jede Katze ein Wirbeltier, aber nicht jedes Wirbeltier eine Katze. Aus diesem Grund werden Hyponymie-Beziehungen auch „ist-ein"-Beziehungen genannt.

2.3.4 *Taxonomien*

Taxonomien entstehen, wenn Hyperonyme und Hyponyme hierarchisch strukturiert sind. Der Begriff „Taxonomie" setzt sich aus den altgriechischen Begriffen „taxis" (Ordnung) und „nómos" (Gesetz) zusammen. In [78] wird eine Taxonomie als „ein einheitliches Verfahren oder Modell definiert, um Objekte eines gewissen Bereichs [...] nach bestimmten Kriterien zu klassifizieren, d.h. sie in bestimmte Kategorien oder Klassen (auch Taxa genannt) einzuordnen". In der Biologie wird eine Taxonomie als Klassenhierarchie verstanden, wie das Beispiel der Unterteilung von Lebewesen in Reiche, Stämme, Klassen Ordnungen, Familien, Gattungen und Arten (siehe Abbildung 3).

10 `http://wordnetweb.princeton.edu/perl/webwn/` - Zugriff am 14.11.2012

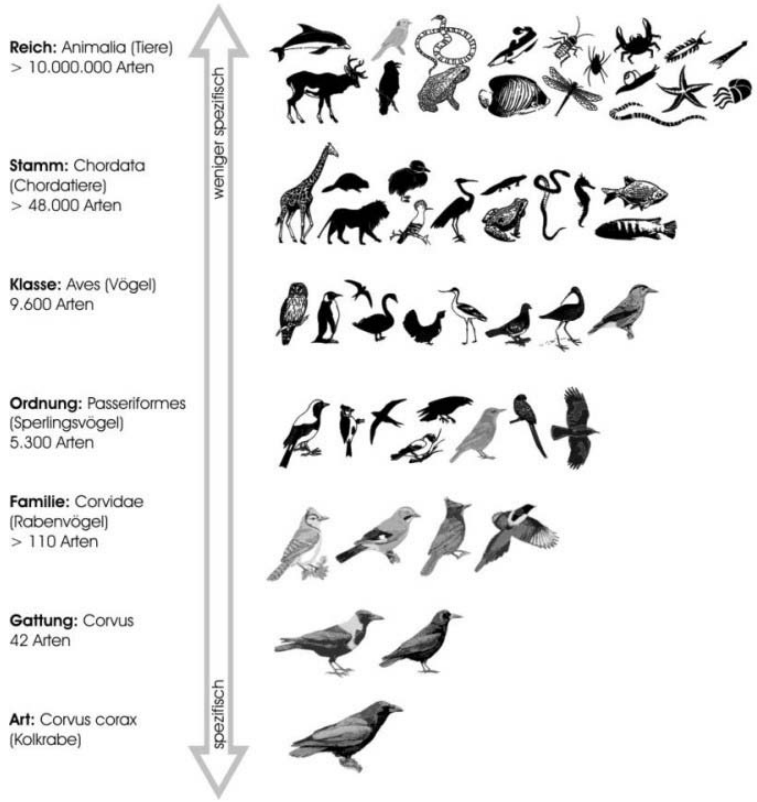

Abbildung 3: Eine Taxonomie in der Biologie [10]

In dieser Arbeit wird die Definition einer Taxonomie an die Definition der Linguistik angelehnt: Eine Taxonomie zeichnet sich durch zwei Eigenschaften aus: Die *Hyponymie* und die *Inkompatibilität* (vgl. [33]). Während die Hyponymie sicherstellt, dass es eine Hyponymie-Relation zwischen Ober- und Unterknoten gibt, besagt die Inkompatibilität, dass Begriffe auf der gleichen hierarchischen Ebene nicht austauschbar sein sollen. Weiter wird in der Linguistik zwischen Klasse-Instanz-Beziehungen, wie z.B. zwischen „Frucht“ und „Apfel“, und reine Hyponymie-Beziehungen, wie z.B. zwischen „Frucht“ und „Pflanze“, unterschieden. Im Rahmen dieser Arbeit werden diese Beziehungstypen unter dem Relationstyp „ist-ein“ zusammengefasst, da diese feinere Unterscheidung für die Anwendungszwecke dieser Arbeit und viele anderer NLP-Anwendungen keine wesentliche Bedeutung hat [152]. Abbildung 4 zeigt ein Beispiel für eine Taxonomie, die verschiedene Fahrzeuge darstellt.

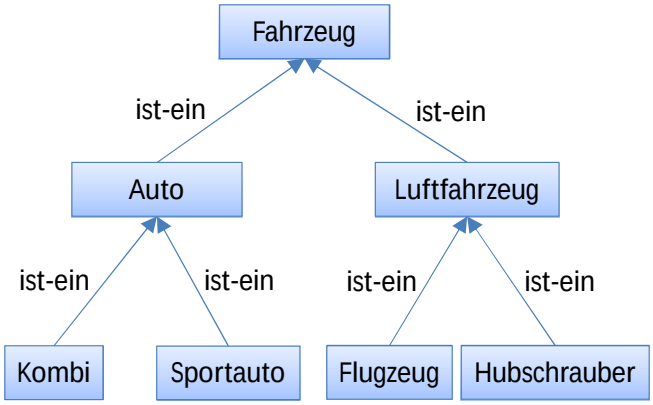

Abbildung 4: Beispiel: Taxonomie

2.3.5 *Thesauri*

Thesauri sind Modelle, die ein Themengebiet repräsentieren. Sie unterscheiden sich von Taxonomien darin, dass neben Hypernymie und Hyponymie weitere Beziehungen zwischen den Konzepten enthalten sind. Die erlaubten Relationen zwischen Beziehungen sind in Normen wie ISO 25964-1[11] definiert. Folgende Relationstypen zwischen Konzepten sind erlaubt:

- Benutzt für (Used for)
- Synonym (Synonym)
- Oberbegriff (Broader term)
- Unterbegriff (Narrower term)
- Verwandter Begriff (Related term)
- Spitzenbegriff (Top term)

Abbildung 5 zeigt ein Beispiel eines Thesaurus, der neben einer kleinen Taxonomie auch einen verwandten Begriff und eine „Benutzt für“-Relation enthält.

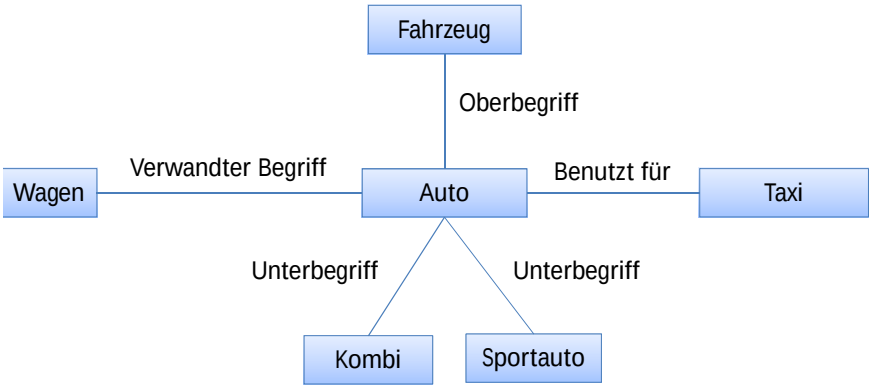

Abbildung 5: Beispiel: Thesaurus

11 http://www.iso.org - Zugriff am 14.11.2012

2.3.6 *Ontologien*

Ontologien sind eine formale und explizite Spezifikation einer gemeinsamen Begriffsbildung [151]. Sie bestehen aus Begriffen und Relationen zwischen diesen Begriffen. Darüber hinaus unterscheiden sie zwischen Begriffen und Instanzen. Instanzen stellen Individuen eines Begriffes dar. Beispielsweise sind „Deutschland“, „Spanien“ und „Italien“ Instanzen des Begriffs „Land“. Relationen zwischen Begriffen können auch auf die Instanzen übertragen werden. Abbildung 6 zeigt eine kleine Beispielontologie mit verschiedenen Relationstypen.

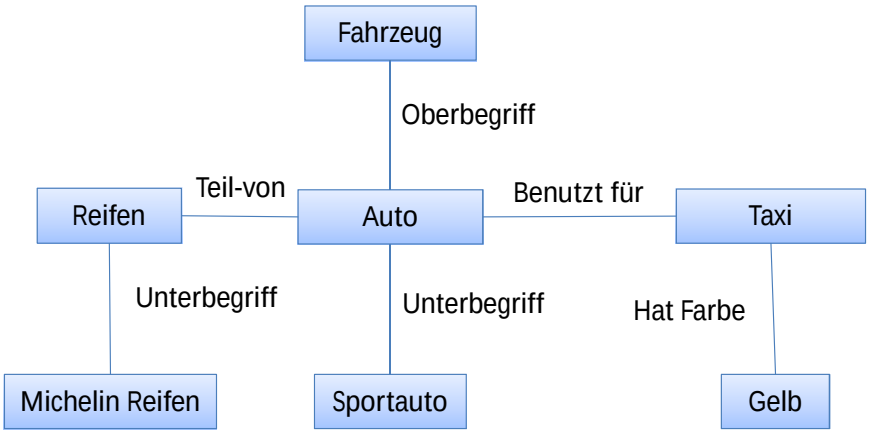

Abbildung 6: Beispiel: Ontologie

Des Weiteren können in einer Ontologie logische Regeln (Axiome) gelten, die die Deduktion von weiteren Regeln ermöglichen. Aus den Fakten „Alle Autos haben Räder“ und „Der VW Touran ist ein Auto“ lässt sich schließen, dass der „VW Touran Räder“ hat. Diese Eigenschaft unterscheidet Ontologien von allen anderen Modellen zur Wissensrepräsentation. Gruber verlangt in [50] außerdem die Maschinenlesbarkeit als eine weitere Eigenschaft von Ontologien. Heutzutage ist das OWL-Format[12] der bekannteste Standard für die Darstellung von Ontologien. Das Beispiel in Abbildung 6 beschreibt die Begriffe „Fahrzeug“ und „Auto“. „Auto“ ist ein Unterbegriff von „Fahrzeug“ und ist definiert als ein „Fahrzeug“ mit dem Wert „car“ im Property „Typ“.

```
<rdf:RDF

  ...

  <owl:Class rdf:ID=''Fahrzeug''/>
  <owl:Class rdf:ID=''Auto''>
    <rdfs:subClassOf rdf:resource="#Fahrzeug"/>
    <owl:equivalentClass>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#Typ"/>
        <owl:hasValue rdf:resource="#auto" rdf:type="#Typ"/>
      </owl:Restriction>
    </owl:equivalentClass>
  </owl:Class>
```

12 http://www.w3.org/TR/owl2-overview - Zugriff am 14.11.2012

```
16    ...

</rdf:RDF>
    \caption{Beispiel: Ontologie-Datei}
                \label{fig:ontology-file}
```

2.3.7 *Semantische Netze*

Semantische Netze haben, genauso wie Taxonomien, keine fest definierten Beziehungstypen. Im Gegensatz zu Ontologien müssen sie nicht formell definiert sein. Sowa hat in [150] ein semantisches Netz als eine graphische Notation zur Darstellung von Wissen definiert. Diese graphische Notation setzt sich aus Knoten, die Konzepte darstellen, und Kanten, die Relationen zwischen ihnen darstellen, zusammen. Jedes Konzept wird durch die Verbindung zu anderen Konzepten definiert. Begriffe und Instanzen werden in semantischen Netzen durch eine spezielle Relation („ist vom Typ“) dargestellt. Beispielsweise gilt: Ein VW Touran „ist vom Typ“ Auto. Zusätzlich dürfen Relationen Unterrelationen haben. Ein Beispiel für ein semantisches Netz wird in Abbildung 7 gezeigt. Die Relation „Haben einen europäischen Hersteller“ hat eine Unterrelation „Haben einen deutschen Hersteller“.

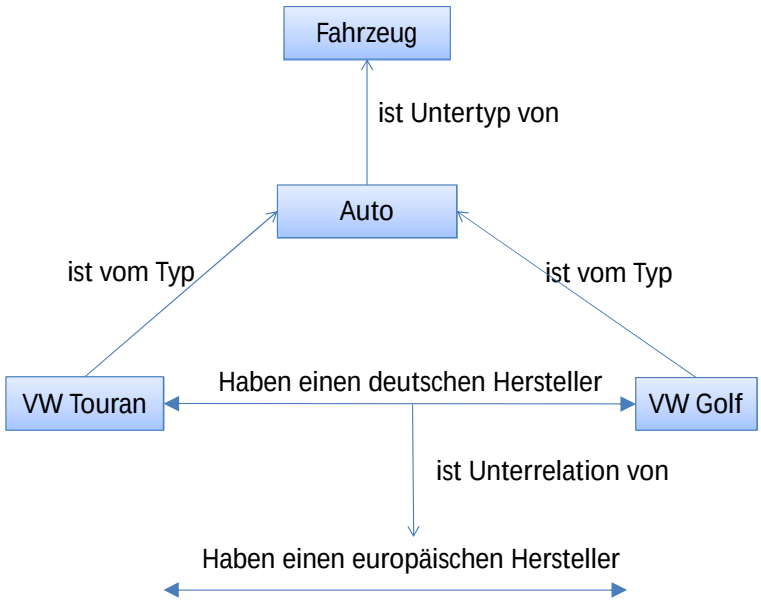

Abbildung 7: Beispiel: Semantisches Netz

2.3.8 *Folksonomien*

Tagging ist der Prozess der Verschlagwortung von *Ressourcen*. Das Schlagwort wird *Tag* genannt. Tagging erfolgt heute in sehr vielen Online-Communities wie delicious[13], Flickr[14] oder YouTube[15].

Eine Folksonomie besteht aus allen Tags, Ressourcen und Benutzern in einer Online-Community. Hotho hat in [61] eine Folksonomie formell als ein 4-Tupel definiert: $F = (U,T,R,Y)$, wobei U die endliche Menge der Benutzer, T die endliche Menge der Tags, R die endliche Menge der Ressourcen in der Folksonomie darstellt. Y ist eine ternäre Relation $Y \in U \times R \times T$, die die Tag-Zuweisungen von Benutzern an Ressourcen repräsentiert. Abbildung 8 zeigt eine kleine Folksonomie, bestehend aus zwei Benutzern, zwei Ressourcen und drei Tags. Ressource 1 wurde mit drei Tags getaggt: „Madrid“, „Weblogs“ und „WWW 2009“. Ressource 2 wurde nur von Benutzer Bob mit dem Tag „Weblogs“ getaggt.

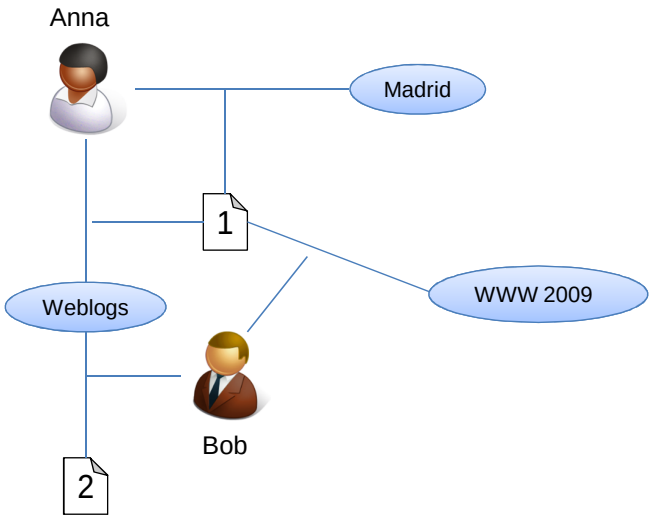

Abbildung 8: Beispiel: Folksonomie

Darüber hinaus wird bei einer Folksonomie die Menge der existierenden Posts, P, betrachtet. Ein Post besteht aus einem Benutzer u, einer Ressource r und allen Tags t_i, mit denen u r getaggt hat. Im oben genannten Beispiel gibt es zwei Posts: $P_1 = (Anna,1,\{„Madrid“,“Weblogs“\})$ und $P_2 = (Bob,2,\{„Weblogs“,“WWW2009“\})$

2.4 WIKIPEDIA

Nachdem im letzten Abschnitt die verschiedenen Möglichkeiten der Wissensrepräsentation dargestellt wurden, sollen im nachfolgenden Abschnitt Wikipedia[16] und ihre Eigenschaften näher betrachtet werden.

13 `http://delicious.com` - Zugriff am 14.11.2012
14 `http://www.flickr.com` - Zugriff am 14.11.2012
15 `http://www.youtube.com` - Zugriff am 14.11.2012
16 `http://www.wikipedia.org` - Zugriff am 14.11.2012

2.4.1 *Das Projekt Wikipedia*

Wikipedia definiert sich selbst als eine „freely lincensed encyclopedia written by thousands of volunteers in many languages“ [160]. Es handelt sich also um eine freilizenzierte Enzyklopädie, die von tausenden Benutzern in vielen Sprachen verfasst wird. Das Wikipedia-Projekt ist im Januar 2001 geboren und entwickelte sich stetig zu einer der zehn populärsten Webseiten[17] der Welt. Seit 2003 wird das Projekt von der Wikimedia Foundation[18] geleitet. Nach dem aktuellen Stand listet die Wikipedia über 23 Millionen Artikel, verfasst von über 1,52 Millionen Autoren[19] in 276 Sprachen[20], auf. Durch die große Anzahl von Freiwilligen kann Wikipedia ein sehr großes Spektrum an Wissensdomänen abdecken. Daraus ergibt sich auch die Tatsache, dass Wikipedia-Artikel sehr aktuell sind. Dieser Fakt sowie die dichte Verweisstruktur [105] machen Wikipedia zu einer attraktiven Quelle für viele Anwendungen [96]. Am Wikipedia-Projekt darf jede Person frei und unentgeltlich teilnehmen, es wird kein Unterschied zwischen Laien, Schülern, Fachleuten oder Forschern gemacht[21]. Inhalte der Wikipedia werden unter der GNU Free Documentation License (GFDL)[22] veröffentlicht, die jedem weitgehende Nutzungsrechte am lizenzierten Werk einräumt. Dadurch lässt sie sich in vielen Projekten einsetzen.

2.4.2 *Struktur der Wikipedia*

Traditionelle Enzyklopädien bestehen aus alphabetisch geordneten Artikeln mit Verweisen zu anderen Artikeln und externer akademischer Literatur. Oft gibt es ein Inhaltsverzeichnis. Wikipedia hat einige dieser Grundsätze übernommen und um wertvolle Elemente ergänzt. Der Aufbau der Wikipedia soll in diesem Abschnitt näher betrachtet werden.

2.4.2.1 *Artikel*

Ein Wikipedia-Artikel beschreibt ein Konzept und bietet deskriptive Texte, Bilder, Listen oder andere Arten von Medien zu diesem Konzept. Ein oder mehrere Begriffe können einem Artikel zugewiesen sein und als Indizes dienen. Beispielsweise leitet die Suche nach „Auto“ in Wikipedia zum Artikel „Automobile“[23] weiter. Ein Ausschnitt dieses Artikels wird in Abb. 9 gezeigt.

17 http://exploredia.com/10-most-visited-websites-2011-2012/ -Zugriff am 14.11.2012
18 http://de.wikipedia.org/wiki/Wikipedia:Sprachen - Zugriff am 14.11.2012
19 http://exploredia.com/10-most-visited-websites-2011-2012/ - Zugriff am 14.11.2012
20 http://de.wikipedia.org/wiki/Wikipedia:Sprachen - Zugriff am 14.11.2012
21 http://de.wikipedia.org/wiki/Wikipedia:Wikipedianer - Zugriff am 14.11.2012
22 http://de.wikipedia.org/wiki/Auto - Zugriff am 14.11.2012
23 http://upload.wikimedia.org/wikipedia/de/1/1d/GNU_Free_Documentation_License_Version_1.2_dreispaltig.pdf - Zugriff am 14.11.2012

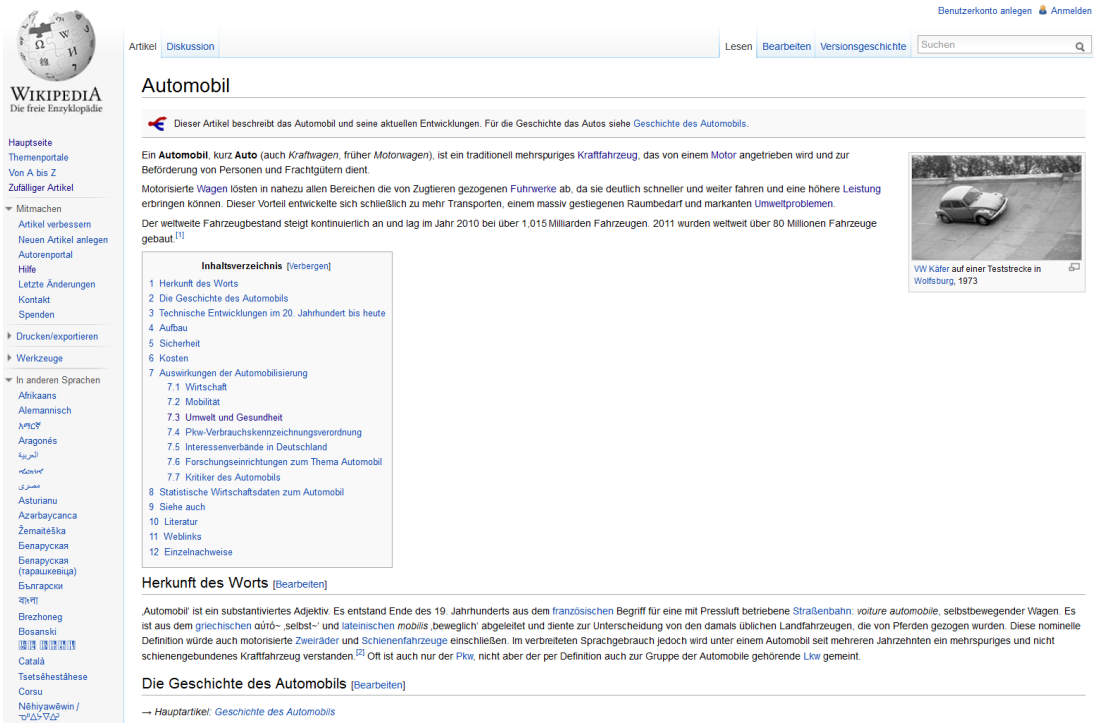

Abbildung 9: Beispiel: Ausschnitt des Wikipedia-Artikels „Automobile“

Des Weiteren verweisen Wikipedia-Artikel zu anderen verwandten Artikeln mittels sogenannter Wikilinks (siehe Abschnitt 2.4.2.2). Sie fangen in der Regel mit einer kurzen Definition an, auch Glosse genannt (siehe 2.4.2.3). Jeder Artikel gehört mindestens einer Kategorie an (siehe 2.4.2.5). Begriffsklärungsseiten (engl. *Disambiguation*) dienen zur Klärung mehrdeutiger Begriffe und verweisen auf die jeweiligen Konzepte (siehe 2.4.2.6)

2.4.2.2 *Wikilinks*

Artikel beschreiben ein Konzept. Diese Beschreibung enthält oft Verweise zu anderen Wikipedia-Artikeln. Beispielsweise erwähnt der Wikipedia-Artikel „Automobile“ andere Konzepte wie „Kraftfahrzeug“[24], „Fuhrwerke“[25] oder „Leistung“[26].

Um die Navigation durch die Wikipedia zu erleichtern, ermöglicht die Wikipedia es, *Wikilinks* zu erstellen. Wikilinks sind Verweise zu verwandten Artikeln. Zusätzlich gibt es noch *Interwikilinks*, die gleiche Artikel in verschiedenen Sprachen miteinander verbinden. Sie können für verschiedene Zwecke benutzt werden, wie die Erkennung von Eigennamen in verschiedenen Sprachen [163], für die Erstellung von parallele Korpora [1] oder multilingualer Wissensquellen [2, 106]. Ferner ist es möglich, einen Wikilink von einem Artikel zu einem bestimmten Abschnitt eines anderen Artikels zu erstellen. Die Verweise zwischen Artikeln lassen sich auf einer allgemeineren Ebene als Graph ansehen, in dem Artikel durch gerichtete Kanten (Wikilinks) miteinander verbunden sind. Dieser Graph wird in der Forschung als *Artikelgraph* bezeichnet.

24 `http://de.wikipedia.org/wiki/Kraftfahrzeug` - Zugriff am 14.11.2012
25 `http://de.wikipedia.org/wiki/Fuhrwerk` - Zugriff am 14.11.2012
26 `http://de.wikipedia.org/wiki/Leistung_%28Physik%29` - Zugriff am 14.11.2012

2.4.2.3 *Glosse*

Die Glosse eines Wikipedia-Artikels wird oft als „Wörterbuch-ähnliche Definition“ charakterisiert [71], die jeder Wikipedia-Artikel besitzen sollte. Laut Wikipedia-Guidelines[27] sollte der erste Paragraph das Konzept definieren. Als Beispiel betrachten wir die Glosse des Wikipedia-Artikel für „Automobile“:

„Ein Automobil, kurz Auto (auch Kraftwagen, früher Motorwagen), ist ein traditionell mehrspuriges Kraftfahrzeug, das von einem Motor angetrieben wird und zur Beförderung von Personen und Frachtgütern dient. Motorisierte Wagen lösten in nahezu allen Bereichen die von Zugtieren gezogenen Fahrwerke ab, da sie deutlich schneller und weiter fahren und eine höhere Leistung erbringen können. Dieser Vorteil entwickelte sich schließlich zu mehr Transporten, einem massiv gestiegenen Raumbedarf und markanten Umweltproblemen. Dies, obwohl der Verbrennungsmotor mitnichten der Antrieb der ersten Stunde war: 1900 verkehrten in den USA 40 Prozent der Automobile mit Dampf, 38 Prozent elektrisch und nur 22 Prozent fuhren mit Benzin. Der weltweite Fahrzeugbestand steigt kontinuierlich an und lag im Jahr 2007 bei rund 918 Millionen Fahrzeugen.“

Abhängig vom Interesse des Lesers kann er sich mit dieser Definition zufrieden geben oder den Artikel weiterlesen, um weitere Details zu erfahren.

2.4.2.4 *Infoboxen*

Infoboxen sind kleine Tabellen, die die Eckpunkte von bestimmten Wikipedia-Artikeln zusammenfassen. Infoboxen gibt es für geographische Einheiten (Kontinente, Länder, Städte, Gemeinden etc), Lebewesen (Pflanzen, Tiere etc) und andere Artikeltypen (Berge, Bands, chemische Elemente etc). Die Infobox des Wikipedia-Artikels „Entenvögel“ wird in Abb. 10 gezeigt. Infoboxen lassen sich aufgrund ihrer maschinenlesbaren Form gut für NLP-Applikationen nutzen, wie das Beispiel DBPedia [11] (siehe 3.2.3) zeigt.

Abbildung 10: Infobox des Wikipedia-Artikels: „Entenvögel“

27 `http://en.wikipedia.org/wiki/Wikipedia:Lead_section` - Zugriff am 14.11.2012

2.4.2.5 *Kategorien*

In Wikipedia gehört jeder Artikel mindestens einer Kategorie an. Die Zugehörigkeit zu einer oder mehreren Kategorien wird mit Hilfe eines Kategorie-Abschnitts am Ende eines Artikels dargestellt, siehe Abbildung 11. Kategorien stellen eine Gruppe von Artikeln zu einem bestimmten Thema dar. Beispielsweise entwählt die Kategorie „Darmstadt“[28] sowohl mit Darmstadt verwandte Artikel wie „Kommunalpolizei Darmstadt“[29], „Luisencenter“[30] oder das „Darmstädter Zentrum für IT-Sicherheit“[31]. Darüber hinaus enthalten sie auch mit Darmstadt verwandte Unterkategorien wie „Bauwerk in Darmstadt“[32], „Stadtteil von Darmstadt“[33] oder „Unternehmen (Darmstadt)“[34]. Unterkategorien dürfen wiederum weitere Artikel oder Kategorien beinhalten.

Abbildung 11: Kategorien-Abschnitt des Wikipedia-Artikels „Darmstadt“

Der Unterschied zwischen Artikeln und Kategorien in Wikipedia kann anhand des Artikels und der Kategorie „Darmstadt“ dargestellt werden. Während der Artikel Darmstadt die Stadt Darmstadt selbst darstellt, ist die Kategorie Darmstadt eine Ansammlung von Artikeln und Unterkategorien, die mit der Stadt Darmstadt verwandt sind. Kategorien und ihre Ober- und Unterkategorien lassen sich als gerichteter Graph (ähnlich wie in Abschnitt 2.4.2.2) darstellen. Der resultierende Graph wird *Kategoriengraph* genannt. Die Kanten im Graph werden durch Kategorienpaare dargestellt, die *Links* genannt werden. *Verfeinerungslinks* (engl. Refinement Links) werden in Wikipedia verwendet, um andere Kategorien zu organisieren [118]. Verfeinerungslinks haben normalerweise die Form „X nach Y“ oder „X als Y“, wobei „X“ und „Y“ beliebige Begriffe darstellen können. Ein Beispiel wäre der Verfeinerungslink „Geographie nach Epoche“. Dieser Link fasst alle Kategorien zusammen, die mit Geographie zu tun haben und strukturiert die hier sich befindenden Kategorien und Artikel nach der Epoche, in der sie stattgefunden haben. Als Beispiel sieht man in Abb. 12 einen Ausschnitt des Kategoriengraphs.

Der Kategoriengraph ist keine reine Taxonomie, da nicht nur Hyponymie-Beziehungen enthalten sind, sondern weitere semantische Relationen. Voess [158] bezeichnet den Kategoriengraph als Thesaurus aufgrund der Tatsache, dass Kategorien zu mehreren Kategorien gehören können. Andere Autoren wie Hammwöhner [52] sehen den Kategoriengraph als strukturiertes Vokabular, mit dem Wikipedia-Artikel verschlagwortet werden können.

28 http://de.wikipedia.org/wiki/Kategorie:Darmstadt - Zugriff am 14.11.2012
29 http://de.wikipedia.org/wiki/Kommunalpolizei_Darmstadt - Zugriff am 14.11.2012
30 http://de.wikipedia.org/wiki/Luisencenter - Zugriff am 14.11.2012
31 http://de.wikipedia.org/wiki/Darmst%C3%A4dter_Zentrum_f%C3%BCr_IT-Sicherheit - Zugriff am 14.11.2012
32 http://de.wikipedia.org/wiki/Kategorie:Bauwerk_in_Darmstadt - Zugriff am 14.11.2012
33 http://de.wikipedia.org/wiki/Kategorie:Stadtteil_von_Darmstadt - Zugriff am 14.11.2012
34 http://de.wikipedia.org/wiki/Kategorie:Unternehmen_%28Darmstadt%29 - Zugriff am 14.11.2012

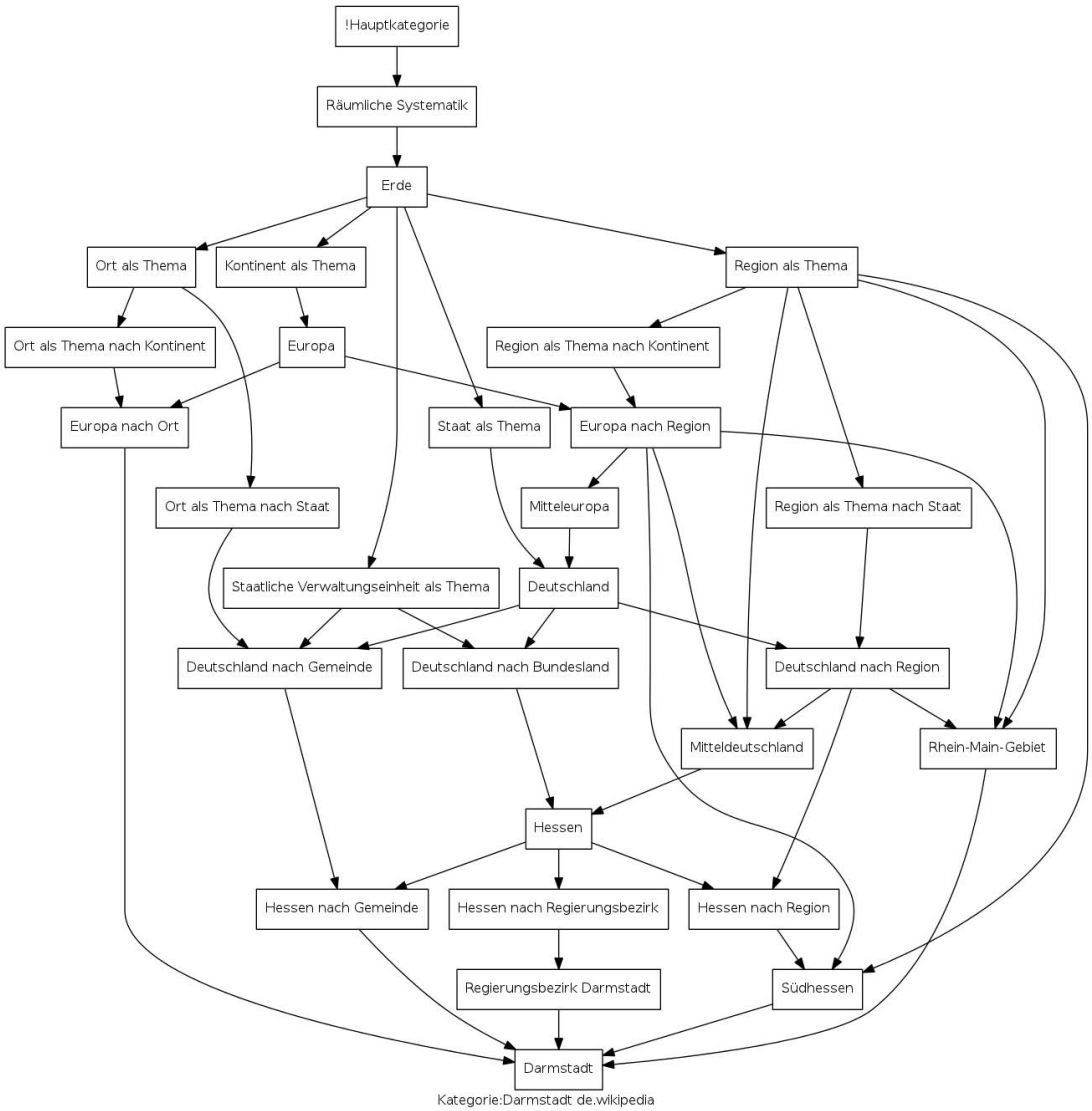

Abbildung 12: Kategoriengraph von der Hauptkategorie zur Kategorie „Darmstadt“

2.4.2.6 *Begriffsklärungsseiten*

Begriffsklärungsseiten (Disambiguierung) werden dazu benutzt, mehrdeutige Begriffe voneinander zu unterscheiden. Wenn ein Benutzer nach einem Wort mit mehreren Bedeutungen sucht, wird er zu einer Begriffsklärungsseite weitergeleitet. Von dieser Seite aus kann der Benutzer weiter zum gewünschten Artikel navigieren. Beispielsweise leitet die Suche nach „Ente“ zu der Begriffsklärungsseite „Ente“[35] (siehe Abb. 13) weiter. Bei Begriffen mit mehreren Bedeutungen wird ein Hinweis auf die Mehrdeutigkeit am oberen Rande des Artikels platziert.

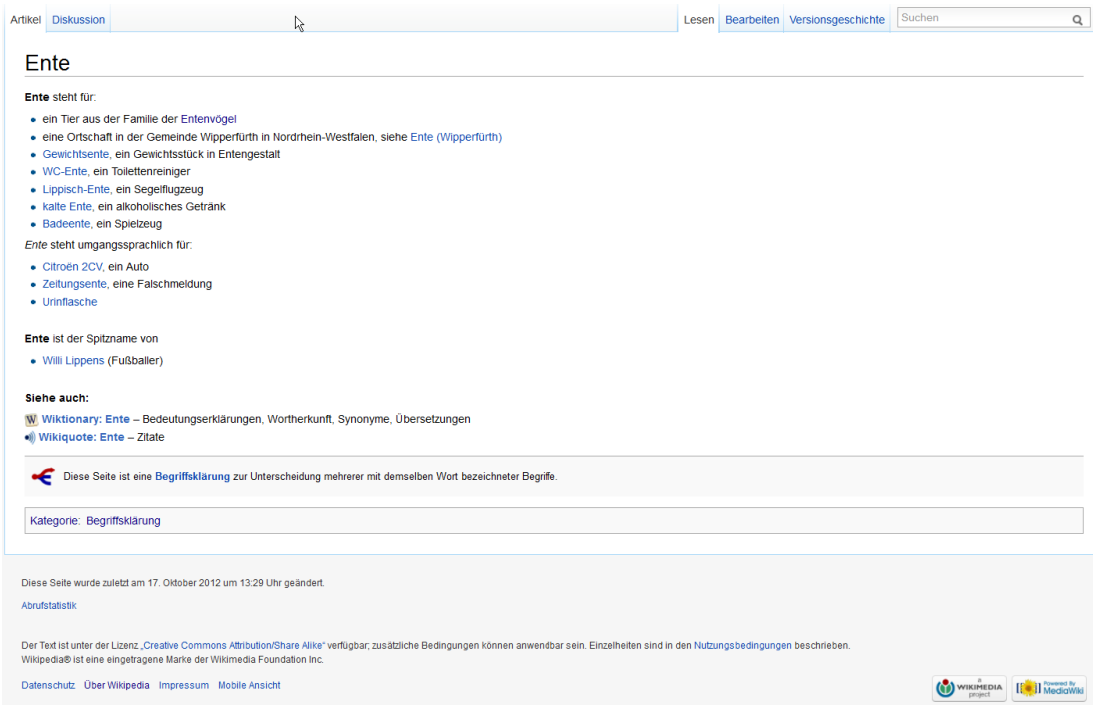

Abbildung 13: Begriffsklärungsseite: Ente

2.4.2.7 *Weiterleitungsseiten*

Weiterleitungsseiten, wie der Name es schon andeutet, leiten von einem Wikipedia-Artikel zu einem anderen. Der Weiterleitungsartikel selbst hat keinen Inhalt, sondern besteht aus einem Wikilink, der zum Ziel-Artikel führt. Beispielsweise gibt es die Weiterleitungsseite „Enten“ (siehe Abb. 14), die zum Artikel „Entenvögel“[36] führt. Weiterleitungsseiten gibt es nicht nur für Pluralseiten, sondern auch für technische Fachbegriffe, Falschschreibungen sowie alternative Schreibweisen.

35 `http://de.wikipedia.org/wiki/Ente` - Zugriff am 14.11.2012
36 `http://de.wikipedia.org/wiki/Entenv%C3%B6gel` - Zugriff am 14.11.2012

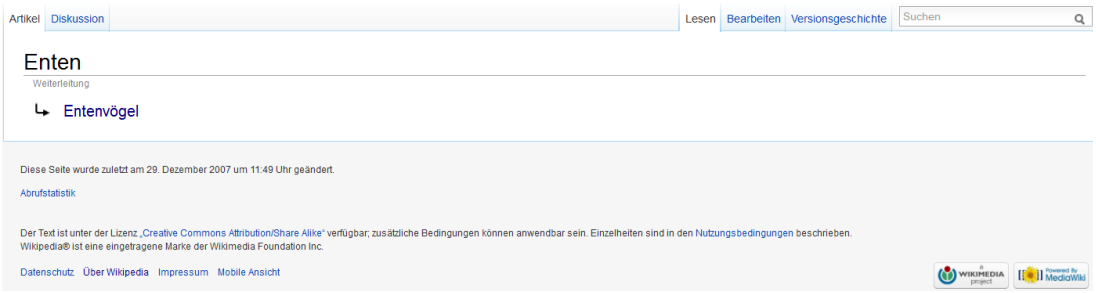

Abbildung 14: Weiterleitungsseite: Ente

2.4.2.8 *Andere Elemente der Wikipedia*

In Wikipedia gibt es außerdem folgende Seiten:

- Hilfeseiten, die Informationen zur Benutzung und Bedienung der Wikipedia enthalten.
- Benutzerseiten[37], auf denen sich registrierte Autoren vorstellen dürfen. Auf Benutzerseiten wird außerdem angegeben, ob ein Benutzer auch Administratorrechte hat.
- Spezialseiten[38], die einen Überblick über den aktuellen Zustand der Wikipedia geben. Beispielsweise definieren diese Seiten, was auf Benutzerseiten stehen darf.
- Vorlageseiten, die vorgefertigte Seiten und Beispiele zur schnellen und einheitlichen Erstellung von Artikeln und Seiten enthalten.

37 `http://de.wikipedia.org/wiki/Spezial:Benutzer` - Zugriff am 14.11.2012
38 `http://de.wikipedia.org/wiki/Hilfe:Spezialseiten` - Zugriff am 14.11.2012

3 VERWANDTE ARBEITEN

»Wir sind gleichsam Zwerge, die auf den Schultern von Riesen sitzen, um mehr und Entfernteres als diese sehen zu können - freilich nicht dank eigener scharfer Sehkraft oder Körpergröße, sondern weil die Größe der Riesen uns zu Hilfe kommt und uns emporhebt.«

— Bernhard von Chartres

Bevor in Kapitel 4 eine Analyse des Anwendungsszenarios des Ressourcenbasierten Lernens in Online Communities erfolgt, die offenen Herausforderungen bestimmt und ein Konzept vorgestellt werden, soll an dieser Stelle ein Überblick über verwandte Arbeiten gegeben werden. Zuerst soll auf Empfehlungssysteme und auf ihren Einsatz im E-Learning eingegangen werden. Danach werden im zweiten Teil verschiedene Ansätze zur automatischen Wissensextraktion behandelt. Dabei liegt der Schwerpunkt auf Verfahren, die Wikipedia als Korpus benutzen, da Wikipedia auch im Rahmen dieser Arbeit für die Wissensextraktion benutzt wird.

3.1 VERWANDTE ARBEITEN IM BEREICH EMPFEHLUNGSSYSTEME

Empfehlungssysteme sind Werkzeuge und Techniken, die dem Benutzer eines Softwaresystems Objekte oder Items vorschlagen [131]. Sie werden benutzt, um Benutzer bei ihren Entscheidungen zu unterstützen wie z.B. beim Kauf von Büchern, bei der Musikauswahl oder der Suche von relevanten Nachrichten im Netz. In diesem Kapitel werden die grundlegenden Ideen und Verfahren von Empfehlungssystemen behandelt.

3.1.1 *Grundlagen zu Empfehlungssystemen*

Seit Mitte der neunziger Jahre [130], als die ersten Beiträge zu Empfehlungssystemen publiziert wurden, gibt es ein wachsendes Interesse an solchen Systemen, das bis heute ungebrochen ist [28]. Insbesondere werden sie in solchen Anwendungen eingesetzt, in denen sehr viele Items zur Verfügung stehen. Das gilt beispielsweise für soziale Netzwerke oder Communities. Empfehlungssysteme zielen darauf ab, Benutzern zu helfen, relevante Items aus einer großen Menge von Items zu finden [27]. Items können Ressourcen aller Art sein: Filme, Lieder, Bücher, Webseiten, Nachrichten, Restaurants, Hotels oder Mode. Heutzutage gibt es aber auch Empfehlungssysteme, die Benutzer [72, 125] oder andere anwendungsspezifische Items wie Tags [69] oder Gruppen [110] empfehlen. Um relevante Empfehlungen für einen gegebenen Benutzer zu berechnen, greifen Empfehlungssysteme zumeist auf die historischen Daten von Benutzern zurück.

Typischerweise wird bei Empfehlungssystemen zwischen den folgenden vier Typen von Ansätzen unterschieden:

- Kollaboratives Filtern

- Inhaltbasiertes Filtern
- Wissensbasiertes Filtern
- Hybrides Filtern

An dieser Stelle soll ein Überblick über die Stärken und die Schwächen der verschiedenen Typen gegeben werden. Anschließend wird ein Blick auf offene Herausforderungen im Zusammenhang mit Empfehlungssystemen geworfen und analysiert, wie sich die verschiedenen Typen für eine Verwendung im E-Learning eignen.

3.1.1.1 *Kollaboratives Filtern*

Diese Art von Empfehlungssystemen stützt sich grundsätzlich auf die Vorlieben der Nutzer, um Listen von Empfehlungen zu generieren. Kollaboratives Filtern [48, 57, 140] erfolgt grob in drei Schritten: Im ersten Schritt werden die Vorlieben und die Präferenzen der verschiedenen Benutzer aus ihrem bisherigen Verhalten identifiziert und anschließend werden sogenannte Nachbarschaften gebildet. Eine Nachbarschaft besteht aus ähnlichen Benutzern (bei Benutzer-basierten Ansätzen) oder ähnlichen Items (bei Item-basierten Ansätzen) [27]. Der letzte Schritt besteht darin, Benutzern Items von anderen Benutzern aus seiner Nachbarschaft anzubieten, die er noch nicht kennt bzw. Items aus der Nachbarschaft der Items des Benutzers. Aufgrund ihrer Einfachheit und Effizienz ist sie eine der verbreitetsten Empfehlungstechniken. Darüber hinaus braucht kollaboratives Filtern keine Informationen über den Inhalt bzw. Bedeutung des Items. Es sind nur die Beziehungen zwischen den Nutzern und Ressourcen von Bedeutung. Nachteile des kollaborativen Filterns sind die Tatsachen, dass eine große Benutzer-Community nötig ist, das *cold-start* Problem [142] für neue Benutzer und Elemente besteht und das *Data-Sparsity* Problem [141]. Das *cold-start* Problem bezeichnet die Situation, wenn ein neuer Nutzer oder eine neue Ressource in die Anwendung hinzukommen. Es lassen sich nicht sofort Empfehlungen generieren, da keine oder wenige Daten über ihn/sie bekannt sind. Auf kollaborativem Filtern basierende Empfehlungssysteme neigen dazu, die beliebtesten Items zu empfehlen, was dazu führt, dass Empfehlungen Richtung Mainstream verfälscht werden. Das *Data-Sparsity* Problem tritt auf, wenn die vorhandenen Informationen nicht ausreichen, um eine geeignete Nachbarschaft eines Benutzers zu bilden.

3.1.1.2 *Inhaltbasiertes Filtern*

Inhaltbasierte Empfehlungssysteme [113] berücksichtigen für die Empfehlungen nur Informationen über die Benutzer und den Inhalt der Ressourcen. Meistens liegen diese Informationen in textueller Form, wie z.B. als Stichworte oder Beschreibungen der Items, vor. Empfehlungssysteme, die auf inhaltbasiertem Filtern aufbauen, suchen automatisch nach Items mit ähnlichen Beschreibungen und empfehlen diese. Diese Art von Empfehlungen hat den Vorteil, dass sie weder auf eine große Benutzer-Community noch auf eine große Profilgeschichte angewiesen ist. Darüber hinaus gibt es das *cold-start* Problem für neue Items nicht. Das *cold-start* Problem für Benutzer besteht dagegen immer noch. Nachteilig ist, dass es zusätzliche Informationen über den Benutzer braucht (i.d.R. die Interessen des Benutzers) und einen Zugriff auf den Inhalt bzw. die Beschreibung der Items (Metadaten). Bei großen Datenmengen erhöht sich die Zeitkomplexität von Empfehlungssystemen aufgrund

der notwendigen Vorverarbeitung und Bearbeitung der textuellen Informationen entsprechend. Beispielsweise müssen u.a. die Wörter im textuellen Inhalt der Items auf ihre Wortstämme zurückgeführt werden, bevor sie mit anderen Wörtern verglichen werden.

3.1.1.3 *Wissensbasiertes Filtern*

Wissensbasiertes Filtern [25, 101] bezeichnet den Prozess der Generierung von Empfehlungen auf Basis einer externen Wissensbasis. Diese Wissensbasen umfassen oft Benutzeranforderungen, Item-Eigenschaften und zusätzliches domänenspezifisches Wissen (z.B. in Form von Taxonomien und Ontologien). Dieses externe Wissen kann vom Empfehlungssystem ausgenutzt werden, um Zusammenhänge zwischen Benutzer oder Items zu erkennen und auf dieser Basis Benutzer oder Items zu empfehlen. Ein Beispiel für diese Art von Empfehlungssystemen ist das sogenannte Ontological Filtering [65]. Damit werden Techniken genannt, die Taxonomien und Ontologien benutzen, um Ähnlichkeiten [89] oder semantische Relationen [104] zwischen Items zu finden. In diesem Zusammenhang spricht man auch über *constraint-based* wissensbasiertes Filtern, worin vordefinierte Empfehlungsregeln bestimmt werden und *case-based* wissensbasiertes Filtern, wo mit Hilfe von Taxonomien und Ontologien und einer Distanzfunktion Ähnlichkeiten zwischen Items berechnet werden. *Constraint-based* wissensbasierte Empfehlungssysteme bestehen aus einem eindeutig definierten Satz von Empfehlungsregeln, die nacheinander angewendet werden [47] und *case-based* wissensbasierte Empfehlungssysteme beinhalten, wie der Name es schon sagt, fallbezogene Regeln, die aus dem vergangenen Verhalten und aus den Wissensbasen abgeleitet werden [86].

Empfehlungssysteme, die auf wissensbasierten Filtern beruhen, eignen sich besonders für die Empfehlung von komplexen Items mit vielen Eigenschaften (z.B. Video-Kameras oder Autos). Vorteile dieser Empfehlungssysteme sind das externe Wissen, das in die Empfehlungen einfließt, die Behebung des *cold-start* Problems und dass Änderungen der Präferenzen eines Nutzers sofort berücksichtigt werden können. Die Nachteile haben meistens mit der benutzten Wissensbasis zu tun: Wie gut passt die Wissensbasis zum Szenario? Was passiert, wenn kein Wissen zu einem Produkt vorhanden ist? Was passiert, wenn Item-Eigenschaften in verschiedenen Sprachen vorliegen?

3.1.1.4 *Hybrides Filtern*

Hybrides Filtern bezeichnet die Kombination verschiedener Datenquellen und Empfehlungssystemansätzen. Dabei unterscheidet man zwischen *parallelisiertem* hybridem Filtern (Empfehlungssysteme werden parallel ausgeführt und anschließend werden die Ergebnisse zusammengeführt), *pipelined* hybridem Filtern (Empfehlungssysteme werden nacheinander ausgeführt, wobei nachfolgende Empfehlungssysteme die Empfehlungsliste der vorherigen Empfehlungssysteme kennen) und monolithischem hybridem Filtern (das monolithische hybride Empfehlungssystem entsteht durch Kombination von Ansätzen und Eigenschaften von verschiedenen Empfehlungssystemen).

3.1.2 *Empfehlungssysteme im E-Learning*

Die ersten Empfehlungssysteme für E-Learning wurden ca. 2003 entwickelt. Es besteht in der Forschung die Übereinkunft, dass sich kommerzielle Empfehlungssysteme nicht einfach auf E-Learning-Systeme transferieren lassen. Tang und McCalla [95, 155, 156] zeigten, dass Empfehlungen, die nur auf Vorlieben der Benutzer beruhen (wie es bei kommerziellen Empfehlungssystemen der Fall ist), oft nicht die pädagogisch wertvollsten sind und dass sich das Ziel, die Rolle und der Kontext eines Benutzers während des Lernens ändern können. Drachsler zählt in [39] verschiedene Kriterien auf, die Empfehlungssysteme in E-Learning berücksichtigen sollten:

- Das Lernziel der Lernenden
- Das Vorwissen der Lernenden (z.B. Anfänger, Fortgeschrittene, Experten)
- Die Eigenschaften der Benutzer
- Erfahrungen von anderen Benutzern
- Lernstrategien von Lernenden

In den letzten Jahren sind aus diesem Grund verschiedene Empfehlungssysteme [39, 93] entwickelt worden, die diesen Anforderungen oder Teilen dieser Anforderungen genügen. Leider sind viele dieser Ansätze nicht über die Entwicklung von Prototypen hinausgekommen und nicht in umfassend genutzte Anwendungen integriert worden. Im Folgenden soll ein Überblick über existierende Systeme gegeben werden.

3.1.2.1 *Auf kollaborativen Filtern basierende Verfahren*

Eines der ersten personalisierten Empfehlungssysteme für E-Learning-Systeme wurde von Recker und Walker [124, 161] entwickelt. Sie verwendeten kollaboratives Filtern und haben untersucht, wie eine Lerncommunity vom Feedback von Lernenden profitieren kann. Tang und McCalla [95, 155, 156] entwickelten ein E-Learning-System, das auf kollaborativen Filtern basierend in der Lage ist, Ressourcen aus dem Web und von spezifischen Anwendungen zu empfehlen. Die Autoren zielten darauf ab, sowohl das Vorwissen als auch die Eigenschaften des Lernenden auszunutzen. Rafaeli et al. [121, 122] haben sich mit der Frage beschäftigt, wie die Zusammenarbeit und die Bildung von Lerngruppen mit Hilfe von kollaborativem Filtern zu fördern ist. Dabei können Lernende entscheiden, ob sie automatisch generierte Empfehlungen oder Empfehlungen von Freunden haben wollen. In [40] hat Dron ein Empfehlungsverfahren entwickelt, das auf kollaborativem Filtern in mehreren Dimensionen besteht. Beispielsweise wird nicht nur berücksichtigt, ob ein Lernender eine Ressource gut gefunden hat, sondern auch andere Eigenschaften wie die Verständlichkeit, die Eignung für Anfänger oder die Art, wie sie geschrieben wurde.

Verfahren, die auf kollaborativem Filtern basieren, eignen sich häufig nicht für E-Learning-Systeme, da sie auf die Verfügbarkeit von großen Datensätzen angewiesen sind. Dies ist im E-Learning meistens nicht der Fall. Einer der Gründe ist die Tatsache, dass viele Systeme für Schulklassen, Lernkurse oder Lehrveranstaltungen mit begrenzten Teilnehmerzahlen konzipiert sind. Darüber hinaus muss ein Lernender viele eigene Ressourcen gefunden haben, bevor Empfehlungen gebildet werden können.

Für das Lernen bedeutet dies, dass gerade beim Beginn einer Recherche bzw. eines Lernprozesses keine Empfehlungen gemacht werden können.

3.1.2.2 *Auf inhaltbasiertem Filtern basierende Verfahren*

Die meisten auf inhaltbasierten Filtern basierende Empfehlungssysteme werden mit anderen Verfahren kombiniert (siehe 3.1.2.4). Dieser Umstand hat mehrere Gründe: Inhaltbasierte Verfahren arbeiten in der Regel daraufhin, ähnliche Ressourcen (vgl. [12, 96, 168]) zu empfehlen. In Anwendungsbereiche, wo die Ähnlichkeit von Objekten eine Rolle spielt, wie z.B. Biotechnologie [114], Geoinformatik [66] oder Linguistik [112], ist dies sehr nützlich. Im E-Learning aber hat dieser Anwendungsfall nur Sinn, wenn Lernende sich für ein spezifisches Thema interessieren, unabhängig von anderen Interessen und Präferenzen. Wenn es aber z.B. darauf ankommt, sich in ein Thema zu vertiefen, reicht die reine Suche nach ähnlichen Ressourcen nicht aus. Schließlich leidet inhaltbasiertes Filtern, genau wie beim kollaborativen Filtern, unter dem *cold-start* Problem.

3.1.2.3 *Auf wissensbasierten Filtern basierende Verfahren*

Auf wissensbasierten Filtern basierende Verfahren lassen sich in drei Kategorien einordnen:

Die erste Kategorie von Verfahren setzt darauf, Kompetenzen von Lernenden und Lernkursen als Ontologie zu modellieren, um auf diese Weise Empfehlungen machen zu können. Zu diesem Verfahrenstyp gehören die Ansätze von Shen et al. [149], Manouselis et al. [92] und Aehnelt [4]. Shen et al. haben basierend auf einer Ontologie und auf Ablaufregeln (engl. *senquencing rules*) ein Verfahren zur Empfehlung von Lernobjekten entwickelt. Das System berechnet die Kompetenz von Lernenden und schlägt Lernobjekte vor, sodass Lernende ihre Kompetenz erhöhen können. Dieses Verfahren setzt eine Modellierung von Kompetenzen voraus, was im Ressourcenbasierten Lernen aufgrund der vielen möglichen Themen und Kompetenzen nicht machbar ist. Manouselis et al. versuchen einen ähnlichen Ansatz. Sie benutzen Ontologien von Lernkursen zu einem Thema, um Lernende durch verwandte Konzepte zu führen. Zusätzlich werden die Präferenzen der Lernenden und die Erfahrungen von ähnlichen Benutzern verwendet. Aehnelt schlägt Empfehlungssysteme für Benutzer vor, die auf eine Modellierung von Lernenden und ihren Kompetenzen beruhen. Dieses Empfehlungssystem berechnet den Bedarf an Wissen von Lernenden auf der Basis von historischen Daten von anderen Lernenden.

Die zweite Kategorie umfasst Ansätze, die darauf abzielen, mögliche Lernpfade zu empfehlen. Dazu gehören die Ansätze von Janssen [68] und Huang et al. [63]. Janssen präsentierte ein Verfahren, das auf der individuellen Lernhistorie beruhend zukünftige Schritte im Lernprozess empfiehlt und Huang et al. versuchen Gruppenlernpfade mit Hilfe von Markov-Ketten zu erkennen. Diese stellen die Wahrscheinlichkeit des Übergangs von einem Lernobjekt zu einem anderen dar. Darüber hinaus wird ein weiteres statistisches Modell benutzt, um neue (unbekannte Pfade) zu erkennen.

Schließlich gibt es die dritte Kategorie von Verfahren, die auf die Modellierung von Lernenden setzt. Khribi et al. [74] greifen auf die Bildung von Benutzerprofilen zurück. Zusätzlich berechnen Sie Ähnlichkeiten zwischen Präferenzen von Lernenden und ihrem Kontext, um hybride Empfehlungslisten (basierend auf kollaborativem Filtern und inhaltbasierten Verfahren) zu erstellen. Jie [70] hat ein personalisiertes

Empfehlungssystem vorgeschlagen. Das System ist in der Lage, abhängig vom Lernstil, Lerntempo und Hintergrundwissen, geeignete Lernmaterialien zu empfehlen. Chen et al. [29] entwickelten ein Empfehlungssystem für Lernkurse, wofür sie Lernprofile bilden (Präferenzen, Interessen und Lernverhalten). Die Empfehlungen hängen dann von den Fähigkeiten der Lernenden ab.

Wie man hier sehen kann, wurden bis heute viele wissensbasierte Verfahren für das Anwendungsszenario E-Learning entwickelt. Diese Ansätze haben gemeinsam, dass sie auf vorgefertigte Ontologien oder Wissen über das Szenario zurückgreifen wie die Modellierung von Kompetenzen oder der Benutzer. Diese Tatsache macht frühere Ansätze für das Ressourcen-basierte Lernen nicht nutzbar, da Ressourcen-basiertes Lernen für das Lernen aller möglichen Themen benutzt werden kann, sodass sich keine Kompetenzen modellieren lassen.

3.1.2.4 *Auf hybridem Filtern basierende Verfahren*

Anderson et al. [5] kombinierten kollaboratives Filtern mit zusätzlichen festen Regeln bzw. Heuristiken, um die Empfehlung von Audio-Lernobjekten zu unterstützen. Die Regeln wurden mit Hilfe einer Domäneontologie definiert und zielen darauf ab, Schwächen des kollaborativen Filterns auszugleichen. Beispielsweise tauchen vertraute Lernobjekte in Empfehlungslisten höher als völlig unbekannte Lernobjekte auf. Koutrika et al. [79] definierten sogenannte flexible Empfehlungen, die mit Hilfe von Operatoren (Filter oder Empfehlungen) gebildet werden können. Abhängig vom Benutzer können inhaltbasierte oder kollaborative Empfehlungen vorgeschlagen werden. Santos [139] schlägt einen hybriden Ansatz vor, der kollaboratives Filtern mit inhaltbasierten Verfahren kombiniert. Dieses Empfehlungssystem setzt Eingaben von Lernenden (Präferenzen, Bewertungen, etc., die das Benutzerprofil bilden) und Lehrenden (Generische Empfehlungen und Annotationen von Ressourcen) voraus. Abhängig von diesen Eingaben des Kontextes (Lernkurs) werden Empfehlungen generiert. Hsu [62] präsentierte ein personalisiertes Online-Empfehlungssystem für die englische Sprache. Empfehlungslisten entstehen aus Kombination von inhaltbasiertem und kollaborativem Filtern zusammen mit weiteren Data-Mining-Techniken. Ziel ist es, geeignete Englischkurse zu empfehlen, in denen Lernende je nach Verhalten in Clustern zusammengefasst werden.

3.1.2.5 *Zusammenfassung*

Im Gegensatz zu den kommerziellen Anwendungen werden rein kollaborative oder rein inhaltbasierte Empfehlungssysteme im E-Learning nur in geringem Maße verwendet. Im E-Learning werden eher wissensbasierte und hybride Empfehlungsverfahren benutzt. Dabei spielt je nach Anwendungsszenario die Modellierung von (Themen-, Lernkurse- oder Kompetenzen-) Ontologien eine große Rolle. Diese ist möglich, weil es sich um geschlossene Szenarien handelt. Dagegen ist das Ressourcen-basierte Lernen sehr offen und kann für die verschiedensten Themenbereiche, Kurse und Lernende benutzt werden. Somit würde eine Modellierung von Kompetenzen, Lernenden oder Kursen nur einem kleinen Teil der Lernenden zugutekommen. Bestehende Verfahren lassen sich im Ressourcen-basierten Lernen daher eher nicht anwenden. Im nachfolgenden Kapitel 4 erfolgt daher eine genaue Analyse des Anwendungsszenarios am Beispiel einer Plattform zur Unterstützung des Ressourcen-basierten Lernens und die in dieser Plattform integrierten Empfehlungssysteme.

3.2 VERWANDTE ARBEITEN IM BEREICH WISSENSEXTRAKTION

Das im Rahmen dieser Arbeit entwickelte Konzept zur Unterstützung des Ressourcenbasierten Lernens in Online-Communities basiert auf einem wissensbasierten Empfehlungssystem, das die mit Hilfe einer Taxonomie generierten zusätzlichen Informationen verwendet, um weitere Items zu empfehlen. Dazu werden in dieser Arbeit zwei Verfahren vorgestellt, die die Taxonomie auf Basis der Wikipedia extrahieren. Die Extraktion von strukturiertem Wissen und dessen Bereitstellung in maschinenlesbarer Form steht im Vordergrund bei vielen Anwendungen aus dem Gebiet des Natural Language Processing. In diesem Abschnitt soll ein Überblick über bestehende Ansätze zur Extraktion von Taxonomien und Ontologien gegeben werden. Zuerst sollen manuell erzeugte Wissensquellen behandelt werden, danach wird auf automatische Verfahren eingegangen. Abschließend werden Ansätze, die auf der Wikipedia basieren, diskutiert.

3.2.1 *Manuell erstellte Wissensbasen*

Die ersten existierenden umfassenden Wissensbasen (Das Wort Wissensbasis wird aus dem Englischen *knowledge base* abgeleitet und bezeichnet alle maschinenlesbaren Wissensquellen wie Taxonomien, Thesauri oder Ontologien) wurden manuell erstellt. Ziel war es, möglichst große Menge an Wissen bei entsprechender hoher Qualität zur Verfügung zu stellen. Einer der populärsten manuell erzeugten Vertreter ist das Projekt *WordNet* [102]. WordNet ist ein semantisches Netz für die englische Sprache. Der Erfolg von WordNet hat gezeigt, dass Wissensbasen für die unterschiedlichsten Anwendungen benutzt werden können. Beispiele für eine Nutzung sind die Sinn-Erkennung von Wörtern [81], die Berechnung der semantischen Ähnlichkeit zwischen Begriffen [24] oder die Sentiment Detection, also die Extraktion von subjektiven Informationen aus Texten [6]. WordNet besteht aus sogenannten *Synsets*. Jedes Synset repräsentiert ein Konzept und besteht aus verschiedenen Wörtern, die die gleiche Bedeutung haben [152]. Beispiele für Synsets sind {Apfelsine,Orange} oder {öffen,aufmachen}. Wie man an diesem Beispiel sieht, können diese „Wörter" nicht nur Substantive, sondern auch Verben, Adjektive und Adverbien sein. Weiter können polyseme Wörter, also Wörter mit mehreren Bedeutungen, wie z.B. „Bank", in mehreren Synsets auftreten. WordNet definiert (abhängig vom Wort-Typ) verschiedene semantische Relationen (siehe 2.3.3) für die Synsets. Ein weiteres Beispiel einer Wissensbasis für die englische Sprache ist Cyc [83]. Cyc verfolgt das Ziel eine umfassende Ontologie des menschlichen Wissens zu erstellen. Im Gegenteil zu WordNet wurde Cyc 1995 von einem Unternehmen (Cycorp[1]) erstellt. Aus diesem Grund gibt es erst seit 2002 *Opencyc*, eine öffentlich verfügbare leicht-abgespeckte Version und seit 2006 *ResearchCyc*, eine für den wissenschaftlichen Einsatz aufbereitete Version. Sie unterscheiden sich darin, dass ResearchCyc sowohl weitere semantische Beziehungen und ein umfangreicheres Lexikon als auch Schnittstellen zur Wissenserweiterung und -bearbeitung zur Verfügung stellt. Cyc besteht aus einer großen Anzahl an einfachen Regeln in Prädikatenlogik, die die verschiedenen Relationen zwischen Konzepten darstellen.

1 http://www.cyc.com/ - Zugriff am 14.11.2012

Schließlich gibt es eine von einer Community manuell erstellte Wissensbasis, Freebase [20]. Freebase wurde von der Firma Metaweb[2] entwickelt und später an Google[3] verkauft. Bei Freebase wird das gesamte Wissen nicht durch Experten erstellt, sondern durch eine Menge von Freiwilligen, ähnlich wie bei Wikipedia. Allerdings gibt es bei Freebase zusätzlich eine strukturierte globale Wissensbasis.

Aufgrund des WordNet-Erfolgs sind in den letzten Jahren ähnliche regionale Projekte zur Erstellung vergleichbarer semantischer Wissensbasen entstanden. Auf der Internet-Seite der Global WordNet Association[4] findet sich ein Verzeichnis mit 69 existierenden Projekten. Viele dieser regionalen Projekte befinden sich leider immer noch in der Entwicklung oder sind nicht frei verfügbar. Für die deutsche Sprache gibt es GermaNet [53]. Dieses hat eine ähnliche Struktur wie WordNet: Neben Wörter-Synsets, die Namen, Verben oder Adjektive sein können, gibt es Relationen zwischen den Synsets. GermaNet enthält außerdem multilinguale Verweise zu EuroWordNet[5]. EuroWordNet ist ein Projekt für europäische Sprachen, das darauf abzielt, ähnliche Wissensbasen wie WordNet für alle europäischen Sprachen zu entwickeln. Darüber hinaus werden die einzelnen Wissensbasen mit Hilfe eines interlingualen Index miteinander verbunden [159].

3.2.2 *Automatische Extraktion von Wissensbasen*

Anstelle einer manuellen Erstellung von Wissensbasen versuchen viele Forscher das Wissen aus existierenden Korpora automatisch zu extrahieren [31]. Sie haben verschiedene Methoden entwickelt, um semantische Relationen zwischen Konzepten zu bestimmen. In der Regel bestehen diese Korpora aus einer Sammlung von Texten, die mit Hilfe verschiedener Methoden verarbeitet werden, um semantische Beziehungen zu erkennen. In diesem Abschnitt sollen diese verschiedenen Methoden und Ansätze gezeigt werden, die zur Erkennung von Hyponymien benutzt werden. Für einen Überblick, wie Ontologien automatisch erstellt und erweitert werden können, wird an dieser Stelle an Faatz [42] verwiesen. Aus der Erkennung von Hyponymien lässt sich eine Taxonomie erzeugen, die im Rahmen dieser Arbeit für das Empfehlungssystem benutzt wird. Die meisten Ansätze zur Erkennung von Hyponymien kombinieren verschiedene Methoden, um die Genauigkeit zu erhöhen, da viele von ihnen Heuristiken sind, die keine 100%ige Genauigkeit anbieten. Aus den Arbeiten von Ponzetto und Strube [118] sowie Merke-Jimenez, Raymond und MacColl [99] lässt sich folgende Klassifizierung der Methoden ableiten:

1. Syntaktische Methoden
2. Lexikalisch-syntaktische Methoden
3. Statistische Methoden
4. Logische Methoden

2 http://en.wikipedia.org/wiki/Metaweb - Zugriff am 14.11.2012
3 http://www.google.de/intl/de/about/ - Zugriff am 14.11.2012
4 http://www.globalwordnet.org/ - Zugriff am 14.11.2012
5 http://www.illc.uva.nl/EuroWordNet/ - Zugriff am 14.11.2012

3.2.2.1 *Syntaktische Methoden*

Bei diesen Methoden ist die Syntax der Wörter für die Bestimmung von Hyponymie-Beziehungen ausschlaggebend. Eine große Rolle bei diesen Methoden spielt der lexikalische Kopf von Lemmata. Das Lemma wird öfters als die Grundform eines Wortes benutzt. Der lexikalische Kopf stellt den Kern einer Phrase dar[6]. Beispielsweise sind die lexikalischen Köpfe der Begriffe „Lernsoftware“ und „Französische Revolution“ die Begriffe „Software“ bzw. „Revolution“. Wie man an diesen Beispielen sieht, weist ein gemeinsamer lexikalischer Kopf auf eine Hyponymie-Beziehung zwischen den beiden Begriffen, die einen gemeinsamen lexikalischen Kopf teilen [118], hin. Die Erkennung des lexikalischen Kopfes einer Phrase wird in der Regel mit Hilfe von externen Werkzeugen wie dem Stanford Parser [75] durchgeführt. In einer Phrase kann der lexikalische Kopf verschiedene Positionen annehmen. So wie in den zwei vorherigen Beispielen der lexikalische Kopf am Ende der Phrase zu finden war, kann er auch am Anfang einer Phrase stehen, wie dies bei „Software für Windows 7“ der Fall ist.

3.2.2.2 *Lexikalisch-syntaktische Methoden*

Die lexikalisch-syntaktischen Methoden sind die meist verbreitetsten Methoden zur Erkennung von Hyponymie-Beziehungen. Das liegt zum einen an der hohen Genauigkeit und zum anderen an ihrer Einfachheit. Lexikalisch-syntaktische Methoden erweitern syntaktische Methoden um eine linguistische Analyse. Lexikalisch-syntaktische Methoden zielen darauf ab, Muster zu erkennen, die auf Hyponymie-Beziehungen schließen lassen bzw. andere Typen von Relationen, um diese auszuschließen. Diese Muster sind heutzutage als Hearst-Patterns [54] bekannt. Die Muster bestehen aus einem festen lexikalischen Bestandteil und aus einem Platzhalter (*NP* steht für Nominalphrase (engl. Noun phrase), d.h. eine Phrase, die einen Namen darstellt) für die aus dem Text zu extrahierenden Begriffe. „NP_0 ist ein NP_1“ und „NP_0 wie NP_1“ sind Beispiele für Hyponymie-Muster, während „NP_0 in NP_1“, „NP_0 of NP_1“ oder „NP_0 hat NP_1“ Beispiele für nicht Hyponymie-Beispiele sind. Wenn in einem Text beispielsweise Phrasen wie „... Ein *Gorilla ist ein Affe*, der ...“ oder „...*Affen* wie *Gorillas* sind sehr stark...“ auftauchen, dann wird eine Hyponymie Beziehung zwischen „Gorilla“ und „Affe“ angenommen. Im Laufe der Jahre wurden die Muster mehrfach modifiziert, erweitert und in weitere Sprachen übertragen [49, 106, 153].

3.2.2.3 *Statistische und maschinelle Lernverfahren*

Maschinelle Lernverfahren analysieren Texte anhand ihrer statistischen Merkmale. In der Forschung wird zwischen überwachtem und unüberwachtem Lernen unterschieden [85]. Beim überwachten Lernen wird ein Verfahren auf einem Trainingkorpus trainiert, der vorausgegangenes Wissen darstellt. Auf dieser Basis kann das Lernverfahren Schlussfolgerungen für weitere Korpora ziehen. Unüberwachte Verfahren lernen direkt aus den Eingabedaten. Die wichtigsten statistischen Verfahren sind das Clustering, die Indexierung mit Hilfe latenter Semantiken und die Kookurrenzanalyse [13, 85]. Die grundsätzliche Idee dieser Verfahren besteht darin, die Häufigkeiten

6 `http://ling.uni-konstanz.de/pages/home/zinsmeister/KL_0910/Material/091118_syntax.pdf` - Zugriff am 14.11.2012

des gemeinsamen Auftretens bestimmter Wörter zur Bestimmung von Informationen über die Art der Relationen zwischen den Wörtern abzuleiten.

Verfahren, die auf Clustering basieren [85, 88, 173], teilen den Dokumentenraum in Teilmengen auf. Dokumente in der gleichen Teilmenge (Cluster) haben eine größere Ähnlichkeit als Dokumente in anderen Mengen. Die Ähnlichkeitsfunktion kann dabei beliebig je nach Ziel gewählt werden. Je nach Ziel kann es sich dann um eine Ähnlichkeitsfunktion handeln, die viele Attribute in Betracht zieht und gewichtet oder nur wenige. Allerdings sagt das Clustern von Begriffen nach ihren Eigenschaften wenig über die Art der Relation zwischen den Begriffen aus. Aus diesem Grund kommen Variationen des Clustering zum Einsatz. Für die Erkennung von hierarchischen Beziehungen kommt z.B. das hierarchische Clustering infrage. Dabei teilt man die verschiedenen Cluster wiederum in kleinere Cluster ein, sodass sich eine Cluster-Hierarchie bildet. Diese Strukturen können zusätzlich als sogenanntes Dendrogramm dargestellt werden (siehe Abb. 16). In der Literatur wird Clustering nicht auf einzelne Terme oder Begriffe angewendet, sondern eher auf Dokumente. Die Benennung der Elemente der Taxonomie erfolgt entweder durch eine Basisontologie oder manuell.

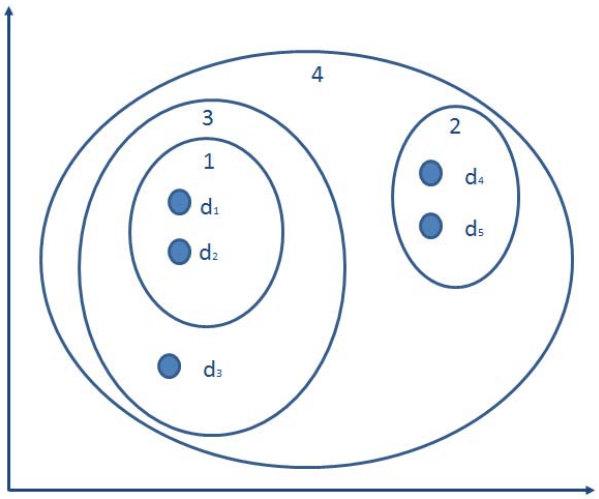

Abbildung 15: Verschiedene hierarchische Cluster

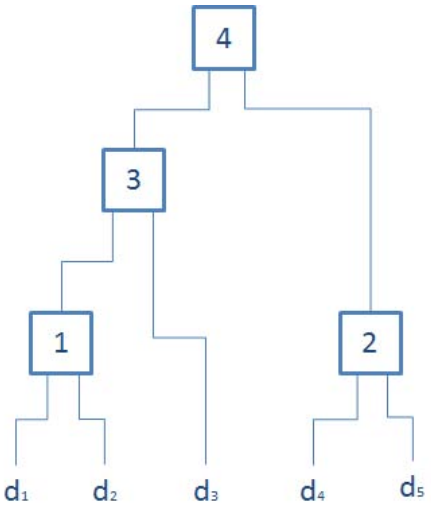

Abbildung 16: Dendrogramm

Latent Semantic Indexing eignet sich für die Suche nach Synonymen in Texten [85]. Es wird angewendet, um statistische Ähnlichkeiten zwischen Begriffen in Texten zu finden. Grundlage für die Berechnung ist eine Term-Dokument-Matrix, mit deren Hilfe Ähnlichkeiten unabhängig von der Schreibweise berechnet werden. Dies wird durch eine geeignete Vorverarbeitung der Terme, z.B. durch Stemming [166], erreicht.

Ein weiteres statistisches Verfahren ist die Kookurrenzanalyse. Diese Analyse sucht in Dokumenten nach semantisch verwandten Termen. Zwei Terme t_1 und t_2 sind genau dann semantisch verwandt, wenn beide Terme statistisch signifikant öfters gleichzeitig als andere Terme auftreten. Beispielsweise werden die Terme „Obama“ und „President“ in einem Nachrichten-Korpus signifikant öfters zusammen auftreten als viele andere Termpaare. Das statistisch signifikante Auftreten von Termen lässt sich mit Hilfe der bedingten Entropie [87] berechnen. Verfahren, die auf Kookurrenzanalyse basieren, können benutzt werden, um Hyponymie-Beziehungen aus

einem Korpus zu extrahieren [138] oder um bereits bestehende Taxonomien oder Hierarchien zu erweitern [137].

3.2.2.4 *Logische Methoden*

Logische Methoden werden hauptsächlich für die Extraktion von Relationen oder Axiomen aus Texten verwendet [13]. Es wird dabei unterschieden zwischen der *induktiven logischen Programmierung* (Aus einer gegebenen Basis werden Konzepte und Relationen abgeleitet) [136, 169] und der *logischen Inferenz* (Erstellung von Relationen mit Hilfe von Transitivitätsregeln und Vererbung) [169]. Da diese Methoden nicht das in dieser Arbeit verfolgte Ziel zum Gegenstand haben, werden sie nur zum Zweck der Vollständigkeit erwähnt.

3.2.3 *Automatische Extraktion von Wissensbasen aus Wikipedia*

Wikipedia ist eine multilinguale, stetig wachsende kollaborative Online-Enzyklopädie mit Abdeckung von einer Vielzahl von Themen. Aus diesem Grund sind in den letzten Jahren verschiedene Verfahren entwickelt worden, die aus Wikipedia semantische Information extrahieren. Zuerst werden in diesem Abschnitt Ansätze erklärt, die nur eine Sprache betrachten, anschließend wird auf multilinguale Ansätze eingegangen.

ISOLDE (Information System for Ontology Learning and Domain Exploration) [162] erzeugt eine Ontologie mit Hilfe von Wikipedia, Wiktionary[7], einem Domain-abhängigen Korpus, einem Identifikator von Eigennamen und einer Basis-Ontologie. Grundsätzlich extrahiert ISOLDE aus dem Domain-abhängigen Korpus Konzept-Kandidaten und nutzt später die Wikipedia, um die Basis-Ontologie mit Hilfe lexikalisch-syntaktischer Methoden zu ergänzen.

Suchanek et al. entwickelten Yet Another Great Ontology (YAGO) [152], die Informationen aus Wikipedia-Kategorien und Infoboxen mit WordNet kombiniert, um eine Domain-unabhängige Ontologie aus Konzepten und Eigennamen zu erstellen. Um dieses Ziel zu erreichen, wird WordNet mit Hilfe von Konzepten aus Wikipedia erweitert. Aus den Infoboxen werden Attribute und Relationen mit Hilfe von Heuristiken extrahiert, wie z.B. die Hauptstadt eines Landes oder der Geburtsort eines Schriftstellers. Aus den Wikipedia-Kategorien werden weitere Relationen, wie z.B. welche Städte in einem Land liegen, abgeleitet. Schließlich werden unplausible oder falsche Daten per Hand gelöscht. Die Evaluation von YAGO weist auf eine Wissensbasis mit einer sehr hohen Qualität bei einer sehr großen Anzahl an Fakten (mehr als 6.000.000) hin.

Wikitology [45] ist eine Wissensbasis bestehend aus Konzepten und Eigennamen. In diesem Ansatz werden strukturierte Daten aus verschiedenen Datenquellen, wie z.B. DBPedia [11] und YAGO, importiert.

DBpedia ist ein Projekt, das auf Community-Arbeit basiert und strukturierte Informationen aus der Wikipedia extrahiert. Hauptquelle von DBpedia sind die Wikipedia-Infoboxen, da diese viele Attribut-Wert-Paare mit spezifischen Informationen zu Wikipedia-Konzepten besitzen. Darüber hinaus enthalten sie semantische Relationen zwischen Individuen. Beispielsweise lässt sich aus der Infobox im Wikipedia-Artikel „Albert Einstein“[8] extrahieren, dass er am 14 März 1879 geboren ist. Herbelot et al.

7 `http://www.wiktionary.org/` - Zugriff am 14.11.2012

8 `http://en.wikipedia.org/wiki/Albert_Einstein` - Zugriff am 14.11.2012

[55] erstellten eine Ontologie basierend auf einem Biologie-Korpus, der aus Wikipedia extrahiert wurde. Chernov et al. [30] konzentrieren sich auf die Unterscheidung zwischen „starken“ und „schwachen“ Relationen bei Wikipedia-Kategorien. Diese Unterscheidung wird aufgrund der Anzahl der Links zwischen Artikeln in den Kategorien durchgeführt: Je mehr Links es gibt, desto stärker ist die Relation.

Ein weiterer Ansatz ist WikiTaxonomy [118]. Es basiert auf Heuristiken und benutzt Wikipedia-Kategorien, um aus Wikipedia und dem Tipster-Sprachkorpus[9] Hyponymierelationen zu extrahieren. Der Tipster-Sprachkorpus ist eine Textkollektion aus Nachrichten, Patenten und wissenschaftlichen Beiträgen und wird benutzt, um die Kategorisierung von Hyponymie-Beziehungen zu verbessern. Der WikiTaxonomy-Ansatz wurde auf die deutsche [73] und die japanische [172] Sprache übertragen. Die Ergebnisse waren aber in beiden Sprachen schlechter, da die Artikel im Deutschen und Japanischen einen geringeren Abdeckungsgrad als im Englischen aufweisen [73, 172]. Sumida et al. [153, 154] extrahieren Hyponymie-Beziehungen aus der Quellcode-Struktur von Wikipedia-Artikeln. Das Ziel von Sumida et al. war die Extraktion einer großen Menge an Hyponymie-Beziehungen aus der japanischen Wikipedia. Yamada et al. [172] erweiterten den Ansatz von Sumida, indem WikiTaxonomy benutzt wurde, um weitere Konzepte zur Menge der Hyponymie-Beziehungen hinzuzufügen.

Ansätze, die die Erstellung multilingualer Wissensbasen mit Hilfe von Wikipedia zum Ziel haben, sind relativ neu und werden erst seit 2011 publiziert. WikiNet nutzt sehr viele Facetten von Wikipedia: Artikel, Wikilinks, Interwikilinks, das Glossar, Infoboxen, Kategorien und den Kategoriengraph. Wikipedia-Artikel und Kategorien stellen die Konzepte dar. WikiNet [106] extrahiert zuerst aus der englischen Wikipedia ein monolinguales semantisches Netz aus Konzepten. Anschließend wird mit Hilfe von Interwikilinks (vgl. 2.4.2.2) ein multilinguales semantisches Netz erstellt, indem die Namen der Wikipedia-Artikel als Konzepte in anderen Sprachen hinzugefügt werden (siehe Abb. 17). Fehlende Interwikilinks können zum Teil ergänzt werden, indem überlappende Interwikilinks benutzt werden [163]. Beispielsweise wenn es keinen direkten Interwikilink zwischen a_1 und a_3 gibt, kann trotzdem ein Interwikilink inferiert werden, wenn es einen Artikel a_2 gibt, der Interwikilinks zu a_1 und a_3 hat. Die Relationen zwischen den Konzepten werden aus dem Kategoriensystem, den Infoboxen sowie dem Artikeltext extrahiert. Die Relationen werden mit Hilfe lexikalischer Methoden ermittelt. WikiNet zeichnet sich durch die hohe Portabilität aus, da als Basis (für den ersten Schritt nicht nur die englische, sondern) eine beliebige Wikipedia-Version gewählt werden kann.

12 “gd”:“Ain-Riaghailteachd” “en”:“Anarchism” “fr”:“Anarchisme” “it”:“Anarchismo” ...
25 “en”:“Autism” “et”:“Autism” “ca”:“Autisme” “fi”:“Autismi” “es”:“Autismo” ...
39 “lt”:“Albedas” “en”:“Albedo” “ast”:“Albedu” “hu”:“Albed” “et”:“Albeedo” ...
290 “lb”:“A (Buschtaf)” “uz”:“A (harf)” “ku”:“A (herf)” “fr”:“A (lettre)” ...
303 “lb”:“Alabama (Bundesstaat)” “br”:“Alabama (stad)” “ro”:“Alabama (stat SUA)” ...
305 “lt”:“Achilas” “fr”:“Achille” “en”:“Achilles” “scn”:“Achilli” “sl”:“Ahil” ...
307 “en”:“Abraham Lincoln” “lv”:“Abrahams Linkolns” “la”:“Abrahamus Lincoln” ...
308 “ga”:“Arastotail” “uz”:“Arastu” “kab”:“Aristot” “fr”:“Aristote” ...
309 “pl”:“Amerykanin w Paryu (Gershwin)” “nl”:“An American in Paris (Gershwin)” ...
316 “en”:“Academy Award for Best Art Direction” “es”:“Anexo:scar a la mejor direccin de arte” ...
324 “en”:“Academy Award” “id”:“Academy Awards” “tr”:“Akademi dlleri” ...

Abbildung 17: Beispiel: Ausschnitt aus WikiNet [106]

9 http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93T3A - Zugriff am 14.11.2012

Navigli und Ponzetto entwickelten mit BabelNet [108] ein multilinguales semantisches Netz durch eine Aggregation von WordNet, Wikipedia und SemCor[10]. SemCor ist ein Text-Korpus, der per Hand mit WordNet-Synsets indiziert wurde. Zusätzlich zu diesen Quellen wurde die Google Übersetzungschnittstelle[11] (Google Translation Application Programming Interface (API)) benutzt, um Wikipedia-Artikelnamen zu übersetzen, für die keine Interwikilinks in anderen Sprachen existieren. In Abbildung 18 wird der Erstellungprozess von BabelNet dargestellt: Aus WordNet extrahiert BabelNet Konzepte sowie alle Relationenstypen, die zwischen ihnen existieren. Im nächsten Schritt werden weitere Konzepte aus Wikipedia-Artikel-Seiten sowie weitere semantische (nicht-spezifizierte) Relationen extrahiert. Interwikilinks werden dann benutzt, um Konzepte in verschiedenen Sprachen zu erkennen. Für fehlende Interwikilinks wird, wie oben erwähnt, der Übersetzungsdienst von Google benutzt. Um die Qualität der Übersetzung zu erhöhen, werden Sätze aus SemCor extrahiert, in denen das gesuchte Konzept vorkommt, übersetzt und die in den übersetzten Sätzen am meisten vorkommende Übersetzung wird dann als richtige Übersetzung des Konzepts angenommen.

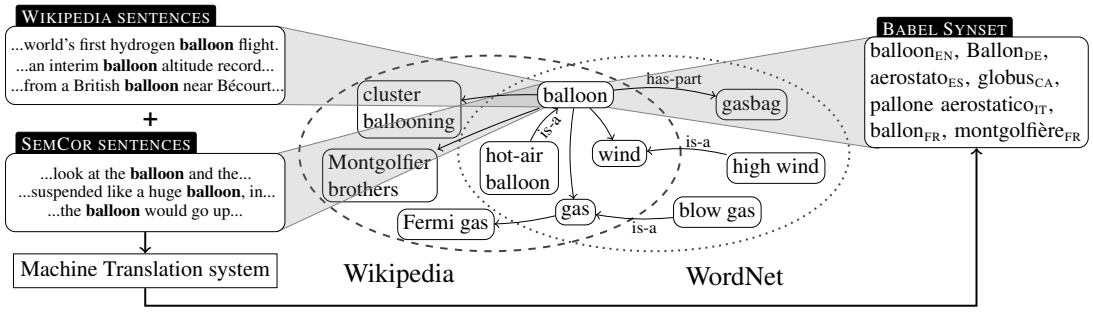

Abbildung 18: BabelNet: Überblick über den Erstellungsprozesses bei BabelNet [108]

Multilingual Entity Taxonomy (MENTA) [98] hat als Ziel die Erstellung einer vollständigen multilingualen Taxonomie, bestehend aus Konzepten und Eigennamen. MENTA versucht dieselben Konzepte in verschiedenen Wikipedia-Sprachversionen nicht nur anhand von Interwikilinks, sondern mit Hilfe von weiteren Heuristiken und syntaktischen Regeln, die manuell spezifiziert werden müssen, zu extrahieren. Im Gegensatz zu WikiNet und BabelNet werden nicht Informationen aus der englischen Wikipedia als Basis für die Erstellung der Wissensbasis verwendet, sondern es werden die Informationen aus allen Wikipedia-Sprachversionen genutzt. Informationen, die in mehreren Wikipedia Sprachversionen vorkommen, werden höher gewichtet als Informationen, die nur in einer Sprache vorkommen. Dieser Prozess wird in Abb. 19 gezeigt: Blau gefärbte Konzepte kommen aus WordNet, gelb gefärbte Konzepte sind Wikipedia-Kategorien und orange gefärbte Konzepte stellen Wikipedia-Artikel dar. Rechts sieht man das Ergebnis des Matchings und der Restrukturierung der Konzepte und Relationen. Für die Bestimmung von Hyponymierelationen zwischen den Konzepten werden syntaktische und strukturelle Eigenschaften von Wikipedia und WordNet in Betracht gezogen. MENTA unterscheidet zwischen zwei Typen von Hyponymierelationen: „Subklasse von“ für die Hyponymie-Relation zwischen zwei Konzepten (bspw. „Hund“ und „Tier“) und „Instanz von“ für die Relation zwischen

10 http://www.cse.unt.edu/~rada/downloads.html#semcor - Zugriff am 14.11.2012
11 https://developers.google.com/translate/ - Zugriff am 14.11.2012

einem Konzept und einer Instanz (bpsw. „Lassie“ und “Hund“). Interwikilinks werden in Abb. 19 rot dargestellt, während die Kanten aus dem Kategoriengraph blau gezeigt werden.

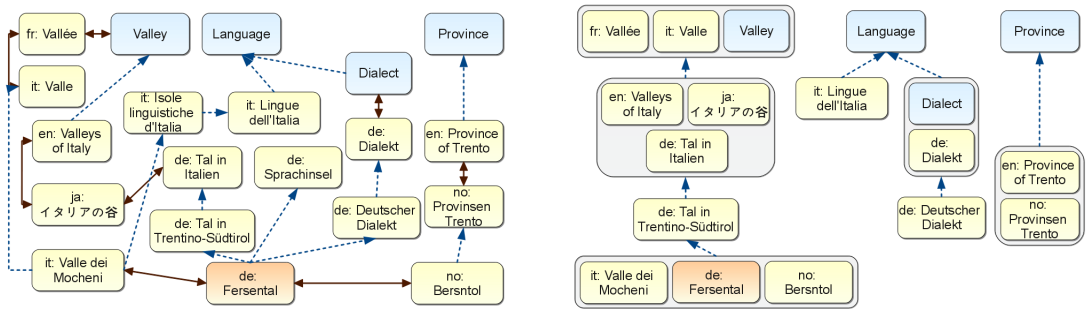

Abbildung 19: MENTA: Vor und nach dem Matching-Prozess von Konzepten aus verschiedenen Sprachen und Quellen [98]

3.2.4 *Diskussion und Einordnung dieser Arbeit*

In den vorherigen Abschnitten wurden verschiedene Verfahren vorgestellt, mit deren Hilfe Wissensbasen aus verschiedenen Quellen gewonnen werden können. Zuerst wurden manuelle Verfahren vorgestellt. Diese Verfahren zeichnen sich durch eine sehr hohe Datenqualität aus, da sie oft durch Experten erstellt werden. Leider bringt die Erstellung durch Experten auch Nachteile mit sich. Zuerst ist die Erstellung durch wenige Personen sowohl zeit- als auch kostenintensiv. Meistens begrenzt sich das Wissen von manuell erstellten Wissensbasen auf domainunabhängiges Wissen, so dass möglichst viele Leute von der Wissensbasis profitieren können. Zudem werden neue Begriffe (z.B. iPad) erst langsam der Wissensbasis hinzugefügt. Seit 2005 wurden z.B. in WordNet nur 3000 neue Konzepte hinzugefügt[12].

Nach den manuellen Verfahren wurden verschiedene automatische (und semiautomatische) Verfahren vorgestellt, um Wissensbasen zu extrahieren oder zu erstellen. Automatische Verfahren, die semantische Relationen aus textuellen Korpora extrahieren, haben sich als nützlich erwiesen, um eine große Menge an Relationen in einer Domain zu erkennen. Die größte Schwäche dieser Verfahren ist die Tatsache, dass sie auf einen „guten“ Korpus angewiesen sind. Um eine allgemeingültige Wissensbasis zu erstellen, müsste man einen Korpus finden, der alle Domains umfasst und realistisch abbildet. In der Realität existiert so ein Korpus nicht, sodass für jede Applikation ein neuer Korpus gebraucht wird. In den letzten Jahren haben Forscher das Potential von Wikipedia als zu verwendendem Korpus erkannt.

Mit Hilfe der im Rahmen dieser Arbeit vorgestellten Ansätze zur Generierung von Wissensbasen unter Nutzung der Wikipedia wurden bereits gute Ergebnisse bzgl. der Genauigkeit und der Abdeckung in anderen Arbeiten erreicht [118, 152, 171]. Allerdings erweisen sich diese Ansätze unflexibel in Bezug auf die Portabilität in andere Sprachen [45, 55, 118, 152–154, 162, 172].

Ansätze, die zusätzlich zu Wikipedia auf andere Wissensbasen wie WordNet oder weitere Sprachkorpora zugreifen, haben den Nachteil, dass sie nicht in andere Sprachen übertragen werden können, da diese Korpora oder Tools oft in anderen Sprachen

12 `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html` - Zugriff am 24.11.2012

nicht verfügbar sind. Darüber hinaus haben die meisten Verfahren die englische Wikipedia-Sprachversion zur Basis. Der Grund dafür ist die Tatsache, dass die englische Wikipedia mit Abstand die größte Wikipedia-Sprachversion ist, sowohl was die Anzahl der Artikel angeht als auch die Anzahl der Autoren. Beispiele dafür sind die multilingualen Ansätze WikiNet und BabelNet. Da aber Wikipedia keine kulturell-neutrale Enzyklopedia ist [115], führt dies dazu, dass *sozio-kulturell spezifisches Wissen*, das in vielen Wikipedia-Sprachversionen enthalten ist, verloren geht. Sozio-kulturell spezifisches Wissen bezeichnet Wissen, das für ein Land, eine Region, für ein Volk oder eine Kultur relevant ist. Diese Art Wissen existiert nicht immer in der englischen Wikipedia. Beispielsweise gibt es in der englischen Wikipedia keinen Artikel vom Widerstand der Anti-Atom-Bewegung in Deutschland zum „Castor-Transport", also dem Transport von radioaktiven Abfällen. In E-Learning-Szenarien, so wie sie im Rahmen dieser Arbeit betrachtet werden, ist dieses Wissen sehr wichtig.

Mehrsprachige Ansätze, die auf Interwikilinks basieren, haben zusätzlich mit dem Problem zu kämpfen, dass Interwikilinks für viele Artikel nicht existieren. Beispielsweise enthalten nur 51,7% aller deutschen Wikipedia-Artikel einen Interwikilink zum jeweiligen englischen Artikel, umgekehrt sind es aber nur 16,6% [143]. Aus diesem Grund arbeitet MENTA mit zusätzlichen Heuristiken zur Erkennung von gleichen Konzepten in verschiedenen Sprachen. MENTA ist ein sehr umfassender und vielversprechender Ansatz, allerdings enthält er viele für die meisten Anwendungsszenarien irrelevante Informationen und ist sehr umfangreich (fast 100 GB). MENTA beinhaltet z.B. eine große Menge an Informationen in der aktuellen Sprache des Benutzers für die keine Übersetzungen vorliegen. Außerdem berichten die Autoren von MENTA, dass für die Erstellung von MENTA oft manuelle Ausnahmen formuliert werden, um die Genauigkeit des Verfahrens zu verbessern [98].

Die Methoden zur Bestimmung von Hyponymierelationen aus den verwandten Arbeiten lassen sich in zwei Gruppen einteilen: Verfahren, die nur auf Wikipedia basieren, und Verfahren, die auch externe Quellen heranziehen. Diese Unterscheidung lässt sich sehr gut am Beispiel des WikiTaxonomy-Ansatzes sehen: Auf der einen Seite gibt es Vorverarbeitungsheuristiken, Heuristiken, die auf Namenskonventionen der Wikipedia basieren und Nachverarbeitungsheuristiken und auf der anderen Seite Syntax- und lexiko-syntaktisch-basierte Methoden, die zusätzlich den Tipster-Corpus benutzen. Aus der Analyse dieser Heuristiken lässt sich schließen, dass externe Quellen herangezogen werden, um das „Verständnis" des Verfahrens zu erhöhen: Während Syntaktische Methoden (vgl. 3.2.3) als sehr akkurat gelten [118, 170], werden lexikalisch-syntaktische Methoden benutzt, um zusätzliche Informationen über Namen, Phrasen und Konzepte zu erhalten. Darüber hinaus sind viele dieser Heuristiken sprachabhängig, da sie die Eigenschaften der englischen Sprache und der Namenskonventionen benutzen, die nicht in jeder Sprache gleich sind.

Aufgrund der dargestellten Schwäche bestehender Verfahren wurden im Rahmen dieser Arbeit zwei Verfahren entwickelt, die die Anforderungen des Szenarios Ressourcen-basierten Lernens erfüllen. Die Anforderungen werden im folgenden Kapitel analysiert und es wird dargestellt, dass die Eigenschaften der Wikipedia (aktuell, mit großer Themenabdeckung, sozio-kulturell spezifisches Wissen und multilingual) gut geeignet sind, die Anforderungen zu erfüllen.

4

UNTERSTÜTZUNG DES KOLLABORATIVEN RESSOURCEN-BASIERTEN LERNENS IN ONLINE COMMUNITIES

»Es ist nicht gut, dass der Mensch alleine sei, und besonders nicht, dass er alleine arbeite; vielmehr bedarf er der Teilnahme und Anregung, wenn etwas gelingen soll.«

— Johann Wolfgang von Goethe

Diese Arbeit hat das Ziel im Ressourcen-basierten Lernen dem Lernenden die Ressourcen, die innerhalb einer Community bereits verwendet wurden, situationsabhängig zugänglich zu machen. In diesem Kapitel soll nun das Anwendungsszenario des Ressourcen-basierten Lernens in Online-Communities konkret vorgestellt werden und auf seine Eigenschaften analysiert werden. Dazu werden unter anderem die CROKODIL-Lernumgebung, eine Plattform zur durchgängigen Unterstützung der mit dem Ressourcen-basierten Lernen verbundenen Aufgaben des Lernenden, und ihr Einsatz betrachtet. Die Analyse zeigt die Schwächen des bisherigen CROKODIL-Ansatzes zur Empfehlung von Ressourcen auf und stellt ein neues auf der Verfügbarkeit einer Taxonomie basierendes Konzept zur Behebung der Schwächen auf. Abschließend werden die Anforderungen an eine zur Realisierung des Konzeptes zu verwendende Taxonomie selbst dargestellt.

4.1 ANALYSE DES ANWENDUNGSSZENARIOS UND DIE CROKODIL-PLATTFORM

Das Anwendungsszenario dieser Arbeit ist das selbstgesteuerte Ressourcen-basierte Lernen. Zur Unterstützung des Ressourcen-basierten Lernens (siehe Kapitel 2.1) wurde die Communities, Web-Ressourcen und Kompetenzentwicklungsdienste integrierende Lernumgebung (CROKODIL)-Plattform [7, 8] entwickelt, die in diesem Kapitel detailliert vorgestellt wird. Eine Betrachtung der Verwendung der CROKODIL-Lernumgebung erlaubt es, Charakteristika des Ressourcen-basierten Lernens zu bestimmen und Schwächen zu identifizieren.

4.1.1 *Ziele der Entwicklung der CROKDODIL-Lernumgebung*

Die CROKODIL-Plattform hat das Ziel, das kollaborative Ressourcen-basierte Lernen mit Web-Ressourcen, wie es in Kapitel 2.1 vorgestellt wurde, zu unterstützen. Die CROKODIL-Plattform will den Herausforderungen, die aus dieser Form des selbstgesteuerten Lernens entstehen, begegnen. Rensing et al. [129] weisen z.B. auf die verschiedenen im Ressourcen-basierten Lernen zu erbringenden Aufgaben hin: Wie in Abbildung 20 gezeigt wird, müssen Lernende nicht nur den aktuellen Informationsbedarf decken, sondern auch andere Aufgaben wie die Planung des Lernprozesses, die Suche nach und die Persistierung von Lernressourcen erfüllen.

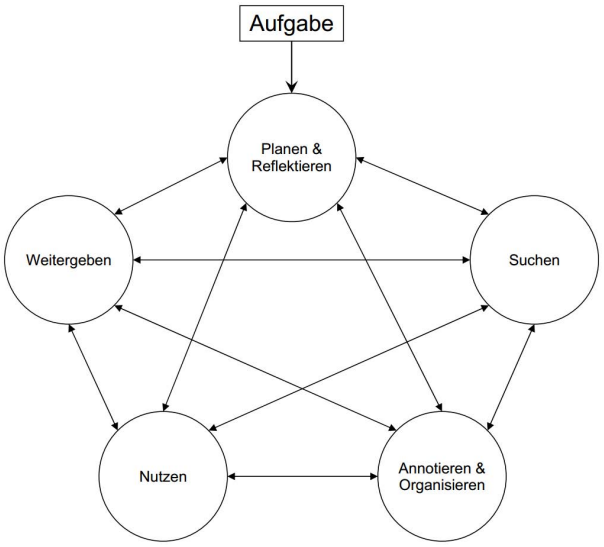

Abbildung 20: Ein Modell zum Ressourcen-basierten Lernen

Für alle diese verschiedenen Aufgaben stehen den Lernenden verschiedene Tools und Programme zur Verfügung:

- Für die Suche werden Web-Browser und Suchmaschinen verwendet.
- Falls die Web-Ressource nicht in HTML-Format vorliegt, braucht der Lernende zusätzliche Programme, um die Datei zu lesen. Bei PDF-Dateien wäre es z.B. ein PDF-Viewer.
- Die Annotation und Organisation von Ressourcen kann z.B. erfolgen mit Hilfe von „Social Bookmarking"-Applikationen wie delicious [1] oder Literaturverwaltungsprogrammen wie JabRef [2].
- Die Weitergabe an Freunde oder Lernpartner könnte u. a. per Mail oder Wikis [147] geschehen.

Es gab bisher keine Applikationen, die den gesamten Prozess des Ressourcen-basierten Lernens bzw. der im Modell genannten Aufgaben unterstützen. Das Ziel der Entwicklung der CROKODIL-Plattform ist es gerade gewesen, alle Schritte des Ressourcen-basierten Lernens zu unterstützen. Da keine andere so umfassende Lernumgebung für Ressourcen-basiertes Lernen bekannt ist, eignet sich CROKODIL im Rahmen dieser Arbeit auch die Anforderungen des Ressourcen-basierten Lernens und die Eigenschaften des Szenarios zu analysieren.

4.1.2 *Funktionalitäten der CROKODIL-Plattform*

Die CROKODIL-Plattform ist eine Web-Applikation, die aus einem Web-Portal und einem Firefox-Plugin besteht. Die CROKODIL-Plattform bietet Funktionalitäten typischer „Social Bookmarking"-Systeme wie delicious, angereichert um Community-Funktionen wie ein Chat- und Nachrichtensystem und Funktionen zur Verwal-

1 http://delicious.com/ - Zugriff am 14.11.2012
2 http://jabref.sourceforge.net - Zugriff am 14.11.2012

tung von Gruppen und Freundschaften. Darüber hinaus ist in der CROKODIL-Plattform das pädagogische Konzept der Aufgabenprototypen (vgl. [127]) implementiert, dass die Selbststeuerung der Lernenden unterstützen soll. Die Realisierung der CROKODIL-Plattform basiert auf semantischen Netzen [94, 150] zur Datenhaltung. Semantische Netze haben sich in den letzten Jahren als ein guter Ansatz zur Unterstützung des Ressourcen-basierten Lernens etabliert, wie Böhnstedt et al. in [17] gezeigt haben. In diesem Abschnitt soll kurz auf die verschiedenen Funktionen der CROKODIL-Plattform eingegangen werden. Abbildung 21 zeigt sie im Überblick.

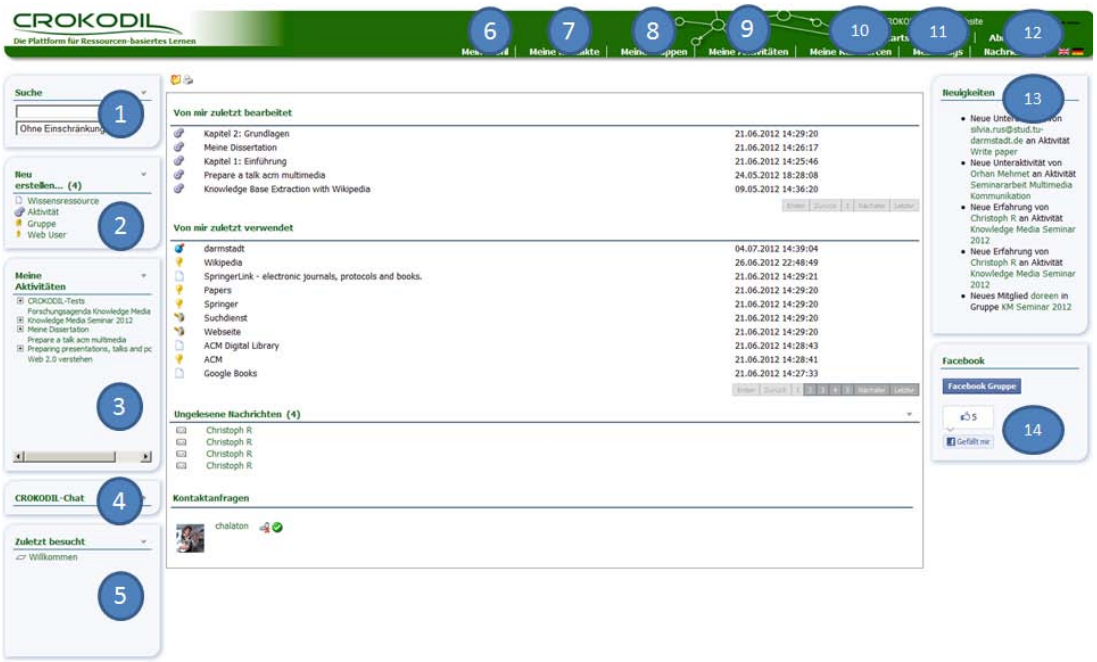

Abbildung 21: Die CROKODIL-Plattform

4.1.2.1 *Planung und Reflektion*

Lernende in der CROKODIL-Plattform können Aktivitäten erstellen (2), die z.B. die Lernziele des Benutzers darstellen [8]. Die Aktivitäten erlauben es dem Benutzer seine Recherchen nach Wissensressourcen vorab zu planen. Eine Aktivität kann außerdem Unteraktivitäten haben (3), die Teilaufgaben in der Bearbeitung einer Aktivität darstellen können. Die Reflektion wird dadurch unterstützt, dass sowohl Aktivitäten als auch Wissensressourcen kommentiert werden können (nicht auf dem Screenshot dargestellt).

4.1.2.2 *Annotierung und Organisation von Ressourcen*

Im Web gefundene Ressourcen, die von den Lernenden im Ressourcen-basierten Lernen genutzt werden, lassen sich in der CROKODIL-Plattform speichern, beschreiben und mit Hilfe von typisierten Tags annotieren, d.h. . es können ihnen Schlagworte zugeordnet werden. Dadurch erhalten Lernende die Möglichkeit, Ressourcen mit semantischer Information anzureichern [18]. Darüber hinaus lassen sich Wissensressourcen zu Aktivitäten zuordnen. Die Zuordnung einer Ressource zu einer Aktivität gibt an, dass die Ressource bei der Bearbeitung zu einer Aktivität verwendet wird.

4.1.2.3 *Suche*

Die CROKODIL-Plattform bietet eine interne Suche (1), die es den Benutzern erlaubt, nach Ressourcen, Aktivitäten, Benutzern, Gruppen oder Tags zu suchen. Darüber hinaus kann sich der Benutzer alle seine Freunde (7), Gruppen (8), Aktivitäten (9), Ressourcen (10) und Tags (11) anzeigen lassen. Tags lassen sich zur Suche und Navigation benutzen.

4.1.2.4 *Community-Funktionalitäten*

Da die CROKODIL-Plattform das Ziel verfolgt, das kollaborative Ressourcen-basierte Lernen, d.h. das Lernen in einer Gruppe, zu unterstützen, bietet es verschiedene Community-Funktionen an:

- Lernende können ein eigenes Profil pflegen (6).
- Es gibt die Möglichkeit, andere Lernende als Kontakte zu speichern (7).
- Lernende können Gruppen bilden (8) und Aktivitäten gemeinsam bearbeiten.
- Lernende können miteinander chatten (4) und sich gegenseitig Nachrichten schicken (12)
- Die Kommentar-Funktion auf der Gruppenseite kann zur Kommunikation zwischen verschiedenen Gruppenmitgliedern benutzt werden.
- Die CROKODIL-Plattform bietet einen Newsfeed mit für den Benutzer potentiell relevanten Informationen über die Aktionen von Lernpartnern oder Kontakten innerhalb der Plattform.

4.1.2.5 *Zugriff auf Informationen und Empfehlungen*

Ressourcen, Aktivitäten und Tags werden in der CROKODIL-Plattform mit anderen Benutzern geteilt. Dazu sind entsprechende Zugriffsrechte an Aktivitäten und Ressourcen festzulegen. Ist dies der Fall, können z.B. alle an einer Aktivität beteiligten Benutzer auf die vom Lernenden zugeordneten Ressourcen zugreifen. Ressourcen, die von Benutzern öffentlich verfügbar gemacht wurden, stehen allen anderen Benutzern zur Verfügung. Sie werden mittels der Suchfunktion gefunden oder können über die zugeordneten Tags gelistet werden. Damit kann ein aktiver Zugriff auf Ressourcen und Informationen anderer Lernender realisiert werden.

Ergänzend werden in CROKODIL Empfehlungen realisiert, mittels derer Lernende aktiv und in Abhängigkeit von ihrer aktuellen Lernaufgabe auf Ressourcen anderer Lernender hingewiesen werden.

CROKODIL verwendet strukturbasierte Empfehlungen auf Basis der Informationen, die im semantischen Netz vorliegen. Bei strukturellen Empfehlungen werden die Kanten zwischen den Knoten traversiert, um zwischenpotenziell interessante, d.h. im Netz in der Nähe befindliche Ressourcen zu finden und den Lernenden vorzuschlagen. Abb. 22 zeigt dafür ein Beispiel. Einem Benutzer, der die Ressource „Semantic Web und E-Learning“ verwendet, könnte die Ressource „Lernen mit Web 2.0“ vorgeschlagen werden, weil beide den gemeinsamen Tag „E-Learning“ verwenden.

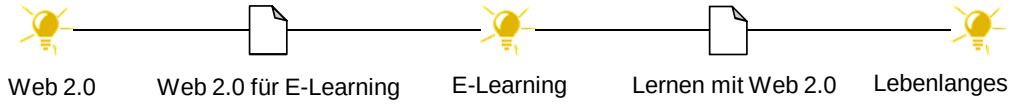

Abbildung 22: Zusammenhängendes semantisches Netz

4.1.3 *Das CROKODIL-Datenmodell*

Das Datenmodell der CROKODIL-Plattform besteht aus verschiedenen Komponenten. In Abbildung 23 wird das Basismodell der CROKODIL-Plattform als Klassendiagramm vorgestellt. Auf die Darstellung der einzelnen Attribute wird an dieser Stelle verzichtet. Das Basismodell deckt sich in weiten Teilen mit dem in [19] ausgearbeiteten Modell. In diesem Kapitel wird auf die für diese Arbeit relevanten Elemente (Ressourcen, Tags und Benutzer) eingegangen und auf die Beschreibung der restlichen Elemente (Gruppen, Aktivitäten, etc) verzichtet.

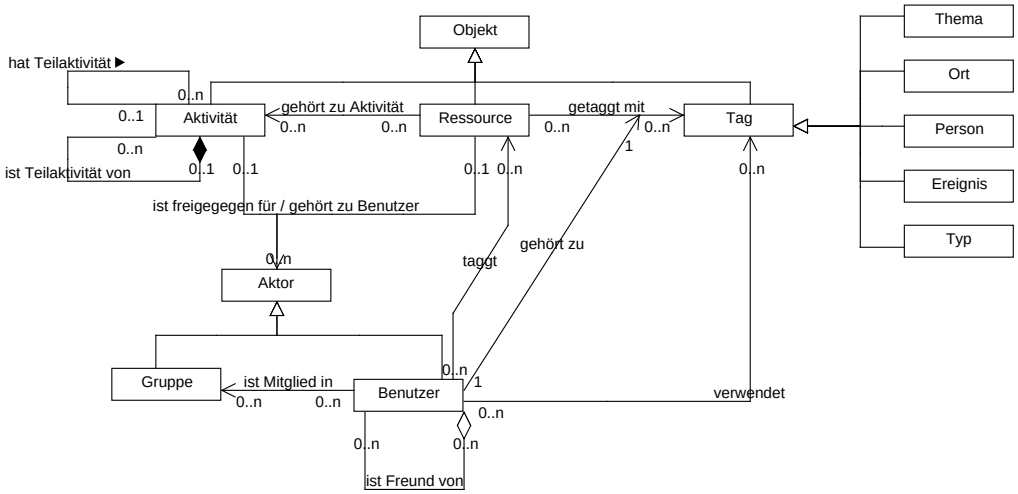

Abbildung 23: Basismodell der CROKODIL-Plattform

Im Modell der CROKODIL-Plattform können Ressourcen gespeichert und ihnen typisierte Tags[18] zugeordnet werden. Typisiertes Tagging verwendet im Gegensatz zu traditionellem Tagging eine anwendungsspezifische Basisontologie von Tagtypen. Mit Hilfe des typisierten Taggings können Benutzer Ressourcen nicht nur mit freien Schlagwörtern taggen, sondern Tags mit eindeutigen Typen wie Personen, Ereignissen, Themen oder Orten verwenden. Im Modell wird weiterhin gespeichert, welcher Aktor ein Objekt (Ressource, Tag oder Aktivität) angelegt hat und ob er Ressourcen mit Tags verschlagwortet hat.

Das CROKODIL-Datenmodell kann als Graph dargestellt werden, wobei Benutzer, Tags und Ressourcen als Knoten repräsentiert werden. Die Kanten stellen die Relationen zwischen den Objekten dar und geben beispielsweise an, dass ein Benutzer eine Ressource gespeichert hat, dass ein Benutzer einen Tag benutzt hat bzw. dass eine Ressource mit einem Tag verschlagwortet wurde. Dies soll anhand folgender Abbildung vereinfacht dargestellt werden:

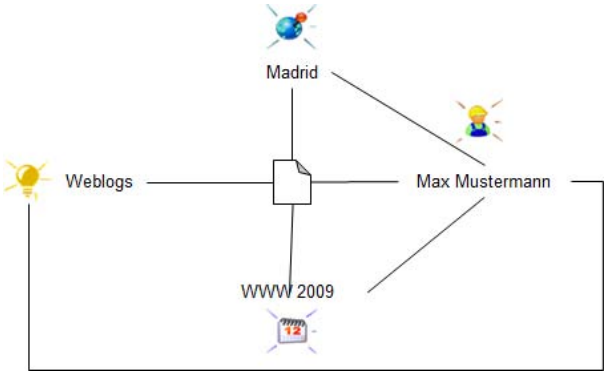

Abbildung 24: Typisiertes Taggen von einer Ressource

In diesem Beispiel hat der Benutzer Max Mustermann eine Ressource (in der Mitte dargestellt) gespeichert und diese mit drei verschiedenen Tags verschlagwortet. Der Tag „Weblogs“ hat den Typ „Thema“, der Tag „Madrid“ den Typ „Ort“ und der Tag „WWW 2009“ den Typ „Ereignis“.

4.1.4 *Analyse der Eigenschaften des Ressourcen-basierten Lernens in Online Communities*

Neben den zuvor bereits angestellten theoretischen Überlegungen zu den Eigenschaften des Ressourcen-basierten Lernens kann die Nutzung der CROKODIL-Plattform durch verschiedene Benutzergruppen und in verschiedenen Szenarien analysiert werden, um die Eigenschaften des Ressourcen-basierten Lernens zu bestimmen. Die CROKODIL-Plattform ist in mehreren Szenarien eingesetzt worden. Diese umfassen sowohl Szenarien, in denen die Lernenden die Plattform ohne Vorgaben individuell nutzen als auch solche, in denen Ressourcen-basiertes Lernen unter einem thematischen Fokus innerhalb von Bildungseinrichtungen [8] stattfindet. In beiden Fällen kann es sich um kurze Episoden, in denen Lernende einen konkreten Wissensbedarf beantwortet oder um längere Episoden, z.B. zur Vorbereitung einer Ausarbeitung zu einem bestimmten Thema, handeln. Innerhalb des Projektes CROKODIL nutzen zwei Bildungseinrichtungen die CROKODIL-Plattform: Das Institut für berufliche Bildung (IBB) und Siemens Professional Education. Diese Bildungseinrichtungen bilden Lernende in verschiedenen Themen und mit verschiedenen Hintergründen aus: So handelt es sich einerseits um Dual-Studierende im kaufmännischen bzw. im elektrotechnischen Bereich und andererseits um Umschüler in IT-Berufen. Zuletzt wurde CROKODIL auch in einer Berufsvorbereitungsmaßnahme eingesetzt. Zudem existiert eine öffentlich zugängliche Instanz der CROKODIL-Plattform, die durch verschiedene Benutzer individuell und durch Studierende des Fachgebietes KOM an der TU Darmstadt genutzt wird. Damit erfolgt eine breite Nutzung der Plattform in sehr heterogenen Szenarien und mit heterogenen Benutzergruppen, so dass die Eigenschaften des Ressourcen-basierten Lernens auf Basis der Szenarien recht umfassend bestimmt werden können.

Die Analyse erfolgte einerseits anhand der Objekte der CROKODIL-Plattform und andererseits mittels der Szenarienbeschreibung der Anwendungspartner im CROKODIL-Projekt. Die Analyse ergab, dass das Ressourcen-basiertes Lernen im CROKODIL-Projekt durch folgende Eigenschaften gekennzeichnet ist: Im Ressourcen-basierten Lernen spielen Ressourcen bzw. Wissen aus verschiedenen Bereichen eine

Rolle. Darüber hinaus spielen im Ressourcen-basierten Lernen aktuelle Themen eine große Rolle, die in Ressourcen im Internet beschrieben sind, aber teilweise nicht in Form von Lehrbüchern vorliegen. Des Weiteren kann man feststellen, dass viele Ressourcen nicht auf Deutsch vorliegen, sondern auf Englisch oder in anderen Sprachen. Beispielsweise ist dies in der Informationstechnologie die Regel. Das bedeutet, dass die CROKODIL-Plattform mit Ressourcen in verschiedenen Sprachen umgehen muss. Auch verwenden die Benutzer zur Verschlagwortung der Ressourcen mittels Tags Begriffe aus verschiedenen Sprachen. Die Analyse der Daten aus der Berufsvorbereitungsmaßnahme ergab weiterhin, dass sozio-kulturell spezifisches Wissen eine Rolle spielt.

Es lassen sich also insgesamt die folgenden Eigenschaften des Ressourcen-basierten Lernens auf Basis der Analyse der umfangreichen CROKODIL-Nutzung festhalten:

- Die Inhalte, die im Ressourcen-basierten Lernen Lerngegenstand sind, sind domänenunabhängig,
- Viele Inhalte, die im Ressourcen-basierten Lernen Lerngegenstand sind, sind sehr aktuell,
- einige Inhalte sind sozio-kulturell spezifisch,
- Inhalte und verwendete Schlagworte sind nicht einsprachig, sondern liegen in unterschiedlichen Sprachen vor, da Informationen unabhängig von der Sprache verarbeitet werden [143].

4.1.5 *Herausforderungen bei der Nutzung von Ressourcen der Community*

Zusammengefasst lässt sich sagen, dass die CROKODIL-Plattform die Herausforderungen des individuellen Ressourcen-basierten Lernens adressiert, indem sie die einzelnen Prozessschritte, wie zuvor dargestellt, unterstützt. Darüber hinaus bietet die Plattform verschiedene Community-Funktionen an, so dass Benutzer in Gruppen und kollaborativ lernen können. Grundsätzlich unterstützt die Plattform damit das Ressourcen-basiertem Lernen in Online-Communities.

Im Ressourcen-basierten Lernen haben die Ressourcen eine zentrale Bedeutung. Wie einführend dargestellt, besteht beim Lernen im Communities die Erwartung, dass Lernende davon profitieren können, wenn sie auf Ressourcen die andere Lernende zum Wissenserwerb genutzt haben hingewiesen werden. In größeren Gruppen oder in einer Community sind für die eigene Lernaufgabe relevante Ressourcen mit hoher Wahrscheinlichkeit bereits von anderen Personen gefunden worden. Diese von anderen Lernenden gefundenen Ressourcen sind in der CROKODIL-Plattform persistiert und mit Schlagworten getaggt. Sie stehen damit den anderen Lernenden grundsätzlich zur Verfügung und es besteht nicht mehr die Notwendigkeit, nach diesen Ressourcen im Web zu suchen.

4.1.5.1 *Analyse der Zugreifbarkeit von Ressourcen anderer Lernender in der CROKODIL-Plattform*

Durch die Community-Funktionen ergeben sich neue Herausforderungen. Im Rahmen dieser Arbeit sollen folgende Herausforderungen adressiert werden:

Betrachtet man die CROKODIL-Plattform und die in der Nutzung der CROKODIL-Plattform gemachten Erfahrungen, so muss man erkennen, dass die Zugreifbarkeit von Ressourcen anderer praktisch eingeschränkt ist. Diese Beobachtung basiert insbesondere auf der Tatsache, dass Lernende bei der Verschlagwortung der Ressourcen unterschiedliche Begriffe verwenden. Dies wird gestützt durch Beobachtungen in [144, 145]. Die Verwendung unterschiedlicher Begriffe hat vielfältige Ursachen:

- Lernende verwenden synonyme Begriffe in einer Sprache, wie beispielsweise „E-Learning“ und „Technology Enhanced Learning“.
- In einem Themengebiet fortgeschrittene Lernende verwenden eher Fachtermini, die Einsteiger in ein Themengebiet noch nicht kennen. Beispielsweise werden in der Biologie von fortgeschrittenen Lernenden lateinische Fachbegriffe verwendet.
- Identisch verwenden fortgeschrittene Lernende eher Spezialisierungen von Begriffen, wohingegen Einsteiger eher generelle Begriffe verwenden
- Lernende verwenden Begriffe in verschiedenen Sprachen.

Die Verwendung unterschiedliche Fachbegriffe führt in CROKODIL dazu, dass die Lernenden Ressourcen anderer Lernender über die Tags nicht finden und zugreifen können. Auch die strukturbasierten Empfehlungen, wie sie in CROKODIL realisiert sind, können mit anderen Schlagworten getaggte Ressourcen nicht empfehlen. Da strukturbasierte Empfehlungen die Kanten zwischen Knoten für die Ähnlichkeitsberechnung ausnutzen, finden Empfehlungen nur statt, wenn es einen Weg zwischen zwei Ressourcen gibt. Wenn es diesen nicht gibt (wie in Abbildung 25), können keine Empfehlungen berechnet werden.

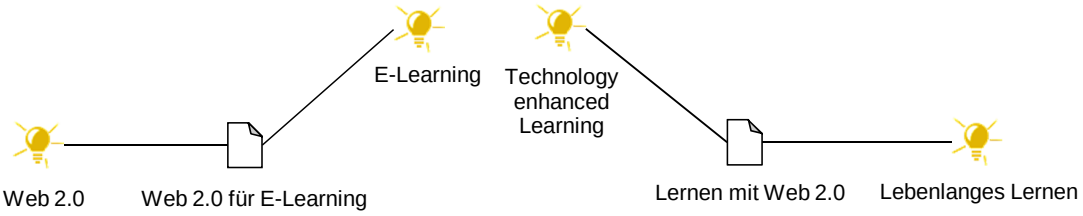

Abbildung 25: Strukturelle Empfehlung ist hier nicht möglich

Diese allgemeine Beobachtung wird gestützt durch die Ergebnisse einer Befragung einer Nutzergruppe der CROKODIL-Plattform innerhalb des Projektes CROKODIL. So wurde die Gruppe der Dual-Studierenden in der Elektrotechnik nach 4 Wochen Arbeit mit der CROKODIL-Plattform befragt, ob ihnen die Empfehlungen aufgefallen sind und ob sie ihnen nützen. Die Ergebnisse sind in den folgenden Abbildungen dargestellt. Sie zeigen, dass offensichtlich eher wenige Empfehlungen in der Plattform angeboten werden, diese aber durchaus von den Lernenden geschätzt werden.

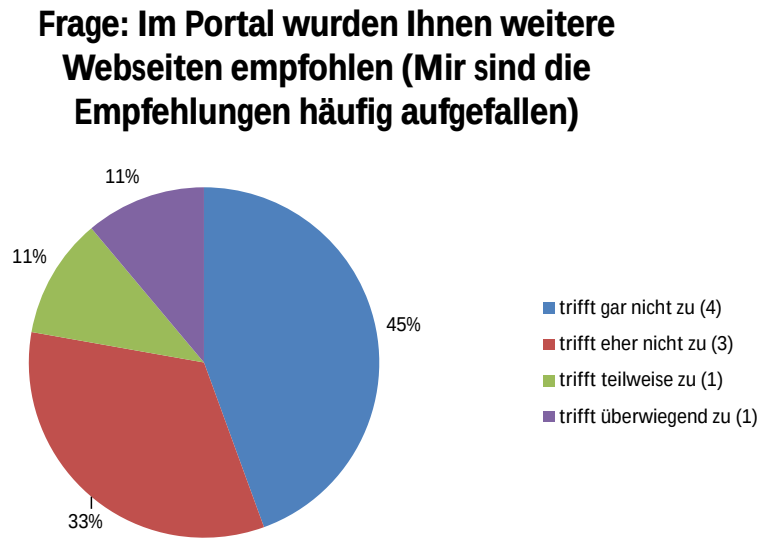

Abbildung 26: Nutzerbefragung: Anzahl der Empfehlungen

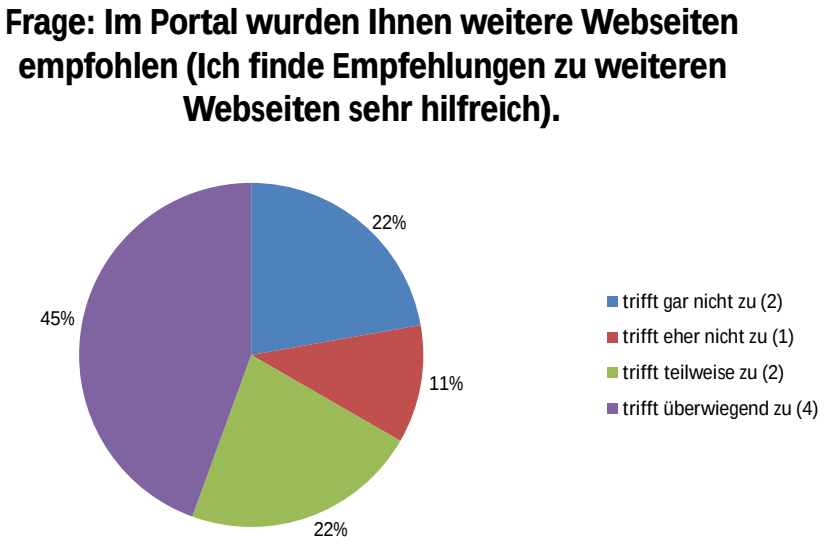

Abbildung 27: Nutzerbefragung: Nützlichkeit der Empfehlungen

4.1.5.2 *Herausforderungen zur Steigerungen der Zugreifbarkeit auf Ressourcen anderer Lernender*

Aus der zuvor vorgestellten Analyse ergeben sich mehrere Herausforderungen, die in dieser Arbeit adressiert werden:

- Wie können alle Lernenden trotz der Verwendung unterschiedliche Begriffe von den in der Community gespeicherten Informationen profitieren, indem sie auf Ressourcen von anderen Lernenden aufmerksam gemacht werden.
- Wie kann die Menge der potentiell relevanten Lernressourcen in den Suchergebnissen erweitert werden, wenn eine Suche keine oder wenige Treffer liefert, indem dem Lernenden zum Beispiel nicht nur Treffer angezeigt werden, die seinen Suchstring enthalten.

- Wie lassen sich hierarchische Strukturen zwischen Themen erkennen, um generelle von spezifischen Ressourcen zu unterscheiden und diese Information den Lernenden zu geben

4.2 KONZEPT ZUR STEIGERUNG DER ZUGREIFBARKEIT AUF RESSOURCEN IM RESSOURCEN-BASIERTEN LERNEN IN ONLINE COMMUNITIES DURCH DIE VERWENDUNG VON TAXONOMIEN

Die Verwendung unterschiedliche Begriffe bei der Verschlagwortung von Ressourcen wurde als eine zentrale Herausforderung analysiert. Das in der Arbeit verfolgte Konzept strebt den Ansatz an, die von den Benutzern verwendeten Begriffe durch Hyponymien zueinander in Beziehung zu setzen. Das bedeutet im konkreten Anwendungsbeispiel CROKODIL, dass die dort vorhandenen Tags mittels neuer Relationen, die die Hyponymiebeziehungen repräsentieren, miteinander verbunden werden. Die Bestimmung der hinzuzufügenden Relationen soll automatisiert, ohne eine manuelle Pflege durch die Lernenden oder einen Administrator der Plattform, erfolgen. Um eine automatisierte Bestimmung der Hyponymiebeziehungen zu realisieren, soll eine Taxonomie verwendet werden. Taxonomien (siehe Abschnitt 2.3.4) wurden als eine Klassifikation von Konzepten in einer geordneten hierarchischen Struktur definiert. Diese Struktur wird typischerweise durch *is-a*-Relationen organisiert. Die *is-a*-Relation ist eine Beziehung zwischen einem Hyperonym und einen Hyponym und besagt, dass das Hyponym „ist-eine (Art)“ Hyperonym. Für alle von den Benutzern verwendeten Tags soll dann paarweise untersucht werden, ob sie sich in der Taxonomie befinden und innerhalb der Taxonomie in einer Hyponymiebeziehung zueinander stehen. Ist dies der Fall, wird eine entsprechende Relation zwischen den Tags im semantischen Netz ergänzt.

Mittels der so ergänzten Relationen lassen sich die oben erwähnten Herausforderungen konkret wie folgt adressieren:

- Durch die ergänzten Hyponymierelationen vergrößert sich der Zusammenhang im semantischen Netz. Auf diese Weise können mehr strukturelle Empfehlungen bereitgestellt werden.
- In der Bestimmung von Suchergebnissen bei einer Suchanfrage durch den Benutzer können ergänzend zu den Treffern einer Volltextsuche Ressourcen angezeigt werden, die mit den gefundenen Ressourcen in hyponymischer Beziehung stehen.
- Die hyponymischen Beziehungen zwischen den Tags und den damit verschlagworteten Ressourcen lassen sich visualisieren, so dass der Benutzer die Zusammenhänge erkennen kann.

Im Folgenden soll dies anhand eines Beispiels dargestellt werden.

4.2.1 *Empfehlung von Ressourcen auf Basis hyponymischer Beziehungen*

Empfehlungen von allgemeinen und speziellen Ressourcen zu einem bestimmten Thema helfen dabei, einen Austausch von Ressourcen zwischen Benutzern zu ermöglichen. Darüber hinaus lassen sich auf diese Weise verschiedene Typen von Lernenden

(abhängig von ihrer Expertise) unterstützen: Anfänger brauchen ggf. mehr generelle Ressourcen, die einen Überblick über das Thema geben, während Experten sich eher evtl. für spezifische Fragen interessieren, sodass sie ihr Wissen vergrößern können [35]. In Abbildung 28 wird ein Beispiel dargestellt: Anna hat zwei Ressourcen mit „Auto“ getaggt, Bob eine mit „Fahrzeug“ und Carl eine mit „Limousine“- Zwischen diesen besteht in der CROKODIL-Plattform keine Verbindung, so dass keine strukturellen Empfehlungen gemacht werden können. Über die Nutzung einer Taxonomie können aber Hyponymie Verbindungen zwischen „Fahrzeug“, „Auto“ und „Limousine“ gefunden werden, so dass Anna sowohl die Ressourcen zu „Fahrzeug“ als auch zu „Limousine“ empfohlen werden können.

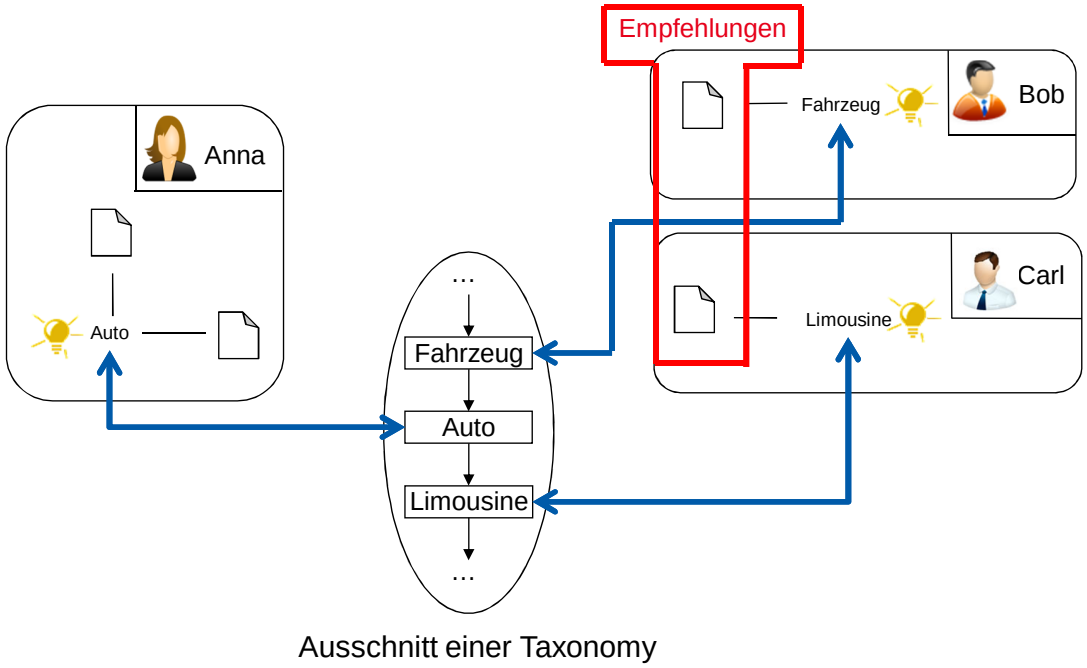

Abbildung 28: Mögliche Ressourcen-Empfehlungen [35]

4.2.2 *Anforderungen an eine Taxonomie zur Ergänzung hyponymischer Beziehungen*

Um das zuvor vorgestellte Konzept zu realisieren, muss eine geeignete Taxonomie verwendet werden. Die Eignung bezieht sich auf das Anwendungsszenario. Dessen Eigenschaften wurden zuvor bereits analysiert. Damit ergeben sich vier Anforderungen an die Taxonomie:

- Domänenunabhängigkeit
- Aktualität
- Abdeckung sozio-kulturell spezifischer Inhalte
- Multilingualität

4.3 ZUSAMMENFASSUNG

In diesem Kapitel wurde die CROKODIL-Plattform als Beispiel für ein System zur Unterstützung des Ressourcen-basierte Lernens in Online-Communities, wie es im Rahmen dieser Arbeit definiert wurde, vorgestellt. Die Analyse der Plattform und ihrer Nutzung ergab, dass Lernende bei der Verschlagwortung von Ressourcen unterschiedliche Begrifflichkeiten verwenden. Diese Tatsache führt dazu, dass Lernende innerhalb einer Community nur eingeschränkt von den Ressourcen anderer Lernender profitieren können.

Aus diesem Grund wurde ein Konzept vorgestellt, das die Lücken im semantischen Netz, die aus der Verwendung unterschiedlicher Begriffe resultiert, mittels der Ergänzung hyponymischer Beziehungen schließen soll. Mit Hilfe dieser ergänzten Relationen soll Abhilfe geschaffen werden, um dem Lernenden einen besseren Zugriff auf Ressourcen anderer Lernender zu ermöglichen.

Das Konzept verlangt die Verfügbarkeit einer Taxonomie deren Eigenschaften aus den Charakteristika des Anwendungsszenarios Ressourcen-basiertes Lernen abgeleitet wurden. Wie in Abschnitt 3.2.3 angemerkt wurde, erfüllt Wikipedia diese Eigenschaften. Die englische Wikipedia deckt mehr als 3 Millionen Konzepte ab, jeden Tag kommen 1000 neue Artikel hinzu und es existieren Wikipedia-Versionen in 281 Sprachen. In den verschiedenen Sprachversionen ist sozio-kulturell spezifisches Wissen enthalten.

In Abschnitt 3.2.4 wurde angesprochen, dass bestehende Ansätze mit Ausnahmen sozio-kulturell spezifisches Wissen nicht oder nicht genügend unterstützen. Ebenso wurde dort MENTA als umfassendster Ansatz analysiert. Die Analyse ergab, dass MENTA für das in Rahmen dieser Arbeit betrachtete Szenario zu „überladen“ ist und auf manuelle Arbeit zurückgreift. Um mit diesen Problemen umzugehen, werden in Rahmen dieser Arbeit zwei Methoden entwickelt und im nachfolgenden Kapitel vorgestellt, die auf einzelne Wikipedia-Versionen unabhängig voneinander angewendet werden können. Da jede Sprache ihre eigene Syntax und Grammatik hat, war das Ziel die Entwicklung eines Ansatzes, der in verschiedenen Sprachen ohne große Änderungen ausgeführt werden kann.

Im Gegensatz zu den monolingualen Ansätzen, die in Abschnitt 3.2.3 erwähnt wurden, benutzen die in dieser Arbeit entwickelten Verfahren nur Wikipedia als Referenzkorpus und basieren nicht auf externen Korpora oder Werkzeugen. Dies erhöht die Sprachportabilität und macht die Verfahren unabhängig von anderen Parteien.

5 ERKENNUNG VON HYPONYMIEN IN VERSCHIEDENEN SPRACHEN

»Taxonomy is described sometimes as a science and sometimes as an art, but really it's a battleground.«

— Bill Bryson

In den vorherigen Kapiteln wurde zum einen ein Überblick über die verschiedenen Methoden und Ansätze gegeben, um Wissensbasen zu erstellen und zum anderen wurde das Konzept zur Unterstützung des Ressourcen-basierten Lernens von Online-Communities auf Basis von Taxonomien vorgestellt. Basierend auf den Anforderungen des Anwendungsszenarios wurden Wikipedia-basierte Ansätze als geeignet für die Umsetzung des Konzepts identifiziert.

In diesem Kapitel werden zwei Verfahren vorgestellt, die ausgehend vom Kategoriengraph der Wikipedia in der Lage sind, Hyponymierelationen im Kategoriengraph zu identifizieren. Als Eingabe wurden für beide Verfahren Kategorienpaare, auch *Links* (vgl. 2.4.2.5) genannt, verwendet. Die Verfahren entscheiden, ob zwischen den im Kategoriengraphen verlinkten Begriffen eine Hyponymierelation existiert oder ob dies nicht der Fall ist. Mit anderen Worten handelt es sich um die Unterscheidung zwischen *is-a*- und *not-is-a*-Relationen.

5.1 ERKENNUNG VON HYPONYMIEN AUF BASIS VON HEURISTIKEN

In Abschnitt 3.2.4 wurde analysiert, dass externe Quellen zur Erkennung von Hyponymierelationen sehr gute Ergebnisse liefern, aber auch Nachteile wurden beschrieben, bspw. in Bezug auf Portabilität in verschiedene Sprachen. TaxWikiHeur.KOM setzt auf die Substitution von sprachabhängigen und auf externen Quellen basierenden Methoden durch andere Heuristiken, die sprachunabhängig sind und allein auf Nutzung der Wikipedia basieren [36].

5.1.1 *Workflow*

Der Ansatz TaxWikiHeur.KOM besteht aus drei Schritten: Ein Vorverarbeitungsschritt, der irrelevante Kategorien und Links eliminiert, ein Hauptschritt bestehend aus vier Heuristiken und ein Nachverarbeitungsschritt, der weitere *is-a*-Relationen (transitiv) propagiert. Der gesamte Workflow wird in Abb. 29 dargestellt.

Die verschiedenen Schritte werden in den nächsten Abschnitten genauer erläutert. Algorithmus 5.1.1 zeigt den TaxWikiHeur.KOM-Algorithmus im Überblick. Die Eingabe des Algorithmus ist eine beliebig große Menge an Links aus dem Kategoriengraph der Wikipedia. Die Ausgabe des Verfahrens ist eine neue Menge von Links, die zusätzlich markiert sind und zeigen, ob zwischen den Kategorien eine Hyponymie-Beziehung existiert. Die Vorverarbeitungsschritte (Zeilen 2-3) verändern

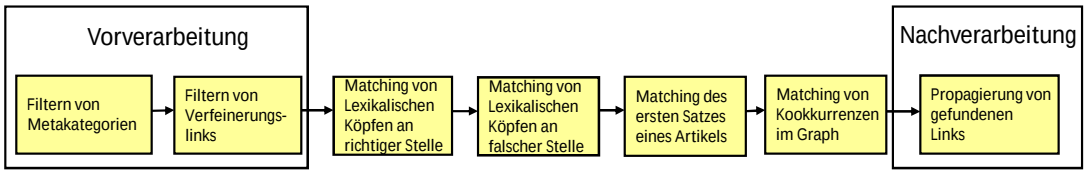

Abbildung 29: Der gesamte Workflow

die Eingabemenge, indem Kategorien und Links eliminiert werden. Die Hauptschritte (Zeilen 4 - 8) bestehen aus vier Heuristiken und tragen die entsprechenden Markierungen an den Links an. Nachverarbeitungsschritte (Zeilen 9 - 10) beenden die Prozedur durch Hinzufügen von transitiven Links.

Algorithmus 5.1.1 TaxWikiHeur.KOM

Eingabe: Eine Menge von Links $L = \{(c_{1,1},c_{1,2}), (c_{2,1},c_{2,2}), \ldots , (c_{n,1},c_{n,2})\}$

1: **Prozedur** TaxWikiHeur.KOM(L)
▷ Vorverarbeitungsschritte
2: Remove_AdminMetacategories(L)
3: Filter_Out_Refinement_Links(L)
▷ Hauptschritte
4: Q = Prepare_Set_For_Labelling(L) ▷ Transformiert alle Paare $(c_{i,1},c_{j,2})$ in markierte Tripel $(c_{i,1},c_{j,2},\varnothing)$
5: Lexical_Head_Matching(Q)
6: Modifier_Matching(Q)
7: First_Sentence_Matching(Q)
8: Cooccurrence_Matching(Q)
▷ Nachverarbeitungsschritte
9: Link_Propagation_Rule(Q)

Ausgabe: Menge von gelabelten Links $Q = \{(c_{1,1},c_{1,2},k_1), (c_{2,1},c_{2,2},k_2), \ldots ,(c_{n,1},c_{n,2},k_n)\}$, wobei $k_i = \{0,1\}$ angibt, ob zwischen beiden Kategorien eine Hyponymie-Beziehung existiert (1) oder nicht (0).

5.1.2 *Einzelne Schritte des Algorithmus im Detail*

In diesem Abschnitt werden die verschiedenen Heuristiken sowie die Vor- und Nachverarbeitungsschritte anhand des Beispiels in Abb. 30 erklärt. Dieses Beispiel zeigt einen kleinen Kategoriengraphen, der einen Ausschnitt aus dem tatsächlichen Kategoriengraph der Wikipedia darstellt, aber aus Gründen der besseren Übersichtlichkeit stark verkürzt wurde. Rechtecke stellen in diesem Graph Wikipedia-Artikel dar, während Wikipedia Kategorien durch Ovale repräsentiert werden.

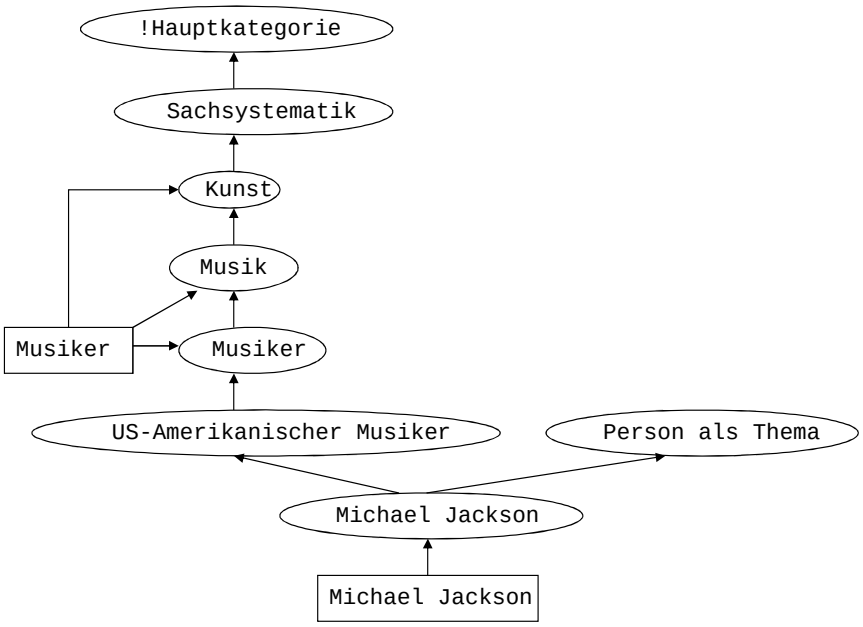

Abbildung 30: Beispiel-Kategoriengraph

Die Menge der Links, also die Eingabe des Verfahrens, lautet wie folgt:

L = {(!Hauptkategorie,Sachsystematik), (Sachsytematik,Kunst),(Kunst, Musik), (Musik,Musiker), (Musiker,US-Amerikanischer Musiker), (Michael Jackson, US-Amerikanischer Musiker), (Michael Jackson,Person als Thema)}

5.1.2.1 *Vorverarbeitungsschritte*

Vorverarbeitungsschritt 1 (Filtern von administrativen Metakategorien)

Der erste Vorverarbeitungsschritt verfolgt das Ziel, den Kategoriengraphen zu säubern, indem Metakategorien, die für administrative Aufgaben benutzt werden, gelöscht werden. Abb. 31 zeigt den exemplarischen Kategoriengraph nach Entfernung der administrativen Metakategorien `!Hauptkategorie` und `Sachsystematik`. In der deutschen Wikipedia kennzeichnen die folgenden Präfixe Metakategorien: `Wikipedia:`, `Wikiprojekte:`, `Artikel:`, `Listen:`, `Kategorien:`, `MediaWiki:`, `Portal:`, `Vorlagen:`, `Hilfe:`, `Sachsystematik`, `Räumliche Sachsystematik`, `Zeitliche Systematik` oder `!` [118]. Dieser Vorverarbeitungsschritt ist vergleichsweise einfach und kann in andere Wikipedia-Versionen transferiert werden. Beispielsweise existiert der Präfix `Kategorien` auch auf Englisch (`Categories`) und Spanisch (`Categorías`).

Algorithmus 5.1.2 Filtern von administrativen Metakategorien in Pseudocode

Eingabe: Eine Menge von Links $L = \{(c_{1,1},c_{1,2}), (c_{2,1},c_{2,2}), \dots , (c_{n,1},c_{m,2})\}$, eine Liste $P = \{p_1, p_2, \dots , p_k\}$ von Präfixen, die Metakategorien im Wikipedia enthalten

1: **Prozedur** REMOVE_ADMINMETACATEGORIES(L,P)
2: **for all** $(c_{i,1},c_{j,2}) \in L$ **do**
3: **if** $lemma(c_{i,1})$ oder $lemma(c_{j,1})$ beginnt mit $p_a \in P$ **then**
4: $L' = L/\{(c_{i,1},c_{j,2})\}$

Ausgabe: Menge von Links $L' = \{(c_{1,1},c_{1,2}), (c_{2,1},c2,2), \dots ,(c_{n,1}, c_{n,2})\}$

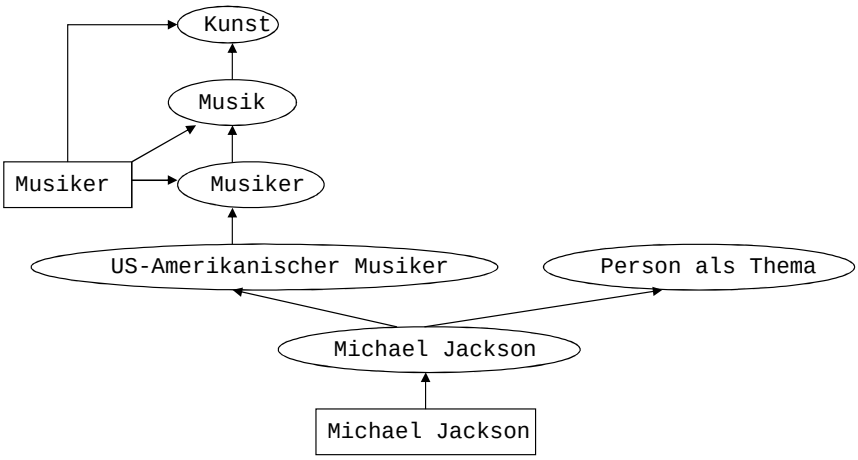

Abbildung 31: Beispiel-Kategoriengraph nach Filterung von administrativen Metakategorien

Vorverarbeitungsschritt 2 (Filtern von Verfeinerungslinks)

Der zweite Vorverarbeitungsschritt hat das Entfernen von Verfeinerungslinks zum Ziel (vgl. Abschnitt 2.4.2.5). Die Heuristik in Pseudocode wird in Algorithmus 5.1.3 angegeben. Der exemplarische Kategoriengraph würde sich auf folgenden Graph reduzieren, denn die Kategorie `Person als Thema` wird entfernt:

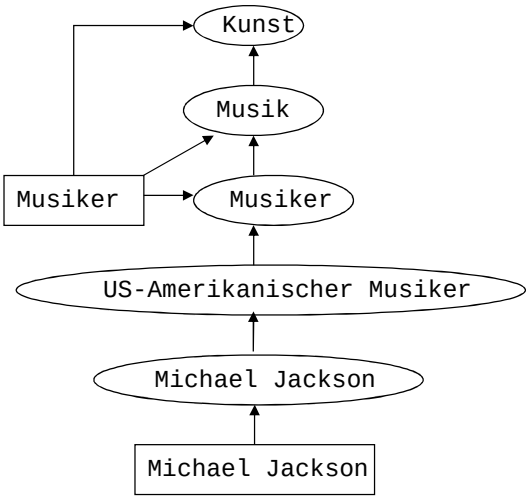

Abbildung 32: Beispiel-Kategoriengraph nach Filtern von Verfeinerungslinks

Die Sprachunabhängigkeit dieser Heuristik wird dadurch gewährleistet, dass in anderen Wikipedia-Sprachversionen auch Verfeinerungslinks existieren. Es muss nur die in dem Verfeinerungslink auftretende Präposition übersetzt werden.

Algorithmus 5.1.3 Filtern von Verfeinerungslinks in Pseudocode

Eingabe: Eine Menge von Links $L = \{(c_{1,1},c_{1,2}), (c_{2,1},c_{2,2}), \dots , (c_{n,1},c_{n,2})\}$, die Präposition p, die in einer gegebenen Sprache zur Angabe von Verfeinerungslinks benutzt wird

1: **Prozedur** FILTER_OUT_REFINEMENT_LINKS(L,p)
2: **for all** $(c_{i,1},c_{j,2}) \in L$ **do**
3: **if** $lemma(c_{i,1})$ or $lemma(c_{j,1})$ enthält String " p " **then**
4: $L' = L/\{(c_{i,1},c_{j,2})\}$

Ausgabe: Menge von Links $L' = \{(c_{1,1},c_{1,2}), (c_{2,1},c_{2,2}), \dots ,(c_{n,1}, c_{n,2})\}$

5.1.2.2 *Heuristiken*

Heuristik 1 (Matching von lexikalischen Köpfen an richtiger Stelle)

Die erste Heuristik basiert auf der in Abschnitt 3.2.2.1 vorgestellten Methode des Vergleichs der lexikalischen Köpfe. Wie bereits erwähnt, kann der lexikalische Kopf eine sehr effektive Methode zur Erkennung von *is-a*-Links [118] sein. Beispielsweise haben „Französische Revolution" und „Revolution" den gleichen lexikalischen Kopf „Revolution". Algorithmus 5.1.4 zeigt den Pseudocode dieser Heuristik. Diese Heuristik muss für die verschiedenen Sprachen adaptiert werden (Zeile 11 in Algorithmus 5.1.4): Bei Kategorien in der englischen Wikipedia ist der lexikalische Kopf in der Regel das letzte Wort, wie bei „Sailboats", „Water sports" oder „Historical reenactment groups".

Algorithmus 5.1.4 Matching von lexikalischen Köpfen an richtiger Stelle in Pseudocode

Eingabe: Eine Menge von ungelabelten Links $L = \{(c_{1,1},c_{1,2},\varnothing), (c_{2,1},c_{2,2},\varnothing), \dots , (c_{n,1},c_{m,2},\varnothing)\}$, eine Liste $P = \{p_1), p_2, \dots , p_k\}$ von Präpositionen in einer gegebenen Sprache

1: **Prozedur** LEXICAL_HEAD_MATCHING(L,P)
2: **for all** $(c_{i,1},c_{j,2},\varnothing) \in L$ **do**
3: **if** $lemma(c_{i,1})$ enthält $p_a \in P$ **then**
4: $c_{i,1} =$ TRUNCATE_PREPOSITIONS$(c_{i,1},p_a)$
5: **if** $lemma(c_{j,1})$ enthält $p_a \in P$ **then**
6: $c_{j,1} =$ TRUNCATE_PREPOSITIONS$(c_{j,1},p_a)$
7: **if** $lemma(c_{i,1})$ enthält Klammern **then**
8: $c_{i,1} =$ TRUNCATE_BRACKETTS$(c_{i,1})$
9: **if** $lemma(c_{j,1})'$ enthält Klammern **then**
10: $c_{j,1} =$ TRUNCATE_BRACKETTS$(c_{j,1})$
11: **if** $lemma(c_{i,1})$ enthält $lemma(c_{j,1})$ an der Position des lexikalischen Kopfes **then**
▷ für Deutsch wird getestet, ob $lemma(c_{i,1}) = w \circ lemma(c_{j,1})$ gilt. $\circ$ stellt die Konkatenation von zwei Zeichen dar und w eine beliebige Zeichenkette
12: $(c_{i,1},c_{j,2},\varnothing) \leftarrow (c_{i,1},c_{j,2},1)$ ▷ Link wird als Hyponymie markiert

Ausgabe: Menge von gelabelten Links $L' = \{(c_{1,1},c_{1,2},1), (c_{2,1},c_{2,2},1), \dots ,(c_{n,1}, c_{n,2},1)\}$

Allerdings gibt es einige Ausnahmen zu dieser Regel, zum Beispiel für Kategorien, die Präpositionen enthalten, wie beispielsweise „Sport in Irland“ und „Campaign for Nuclear Disarmament“ oder jene Kategorien, die eine Kategorie mit Hilfe eines Begriffs in Klammern verfeinern wie beispielsweise „Sport (Irland)“. Dieses Problem kann für die meisten Fällen dadurch gelöst werden, dass diese Ausnahmen explizit definiert werden und die restliche Phrase nach der Präposition ignoriert wird oder Begriffe in der Klammer nicht betrachtet werden (Zeilen 3 - 10 in Algorithmus 5.1.4). Die Wikipedia-Nomenklatur für die Verfeinerung von Kategorien mit Präpositionen und Begriffen in Klammern existieren in anderen Sprachen auch und können dementsprechend genauso in anderen Sprachen behandelt werden.

Im exemplarischen Kategoriengraph bedeutet dies, dass der Link (`Musiker`,`US-amerikanischer Musiker`) als Hyponymie erkannt wird. Die Markierung dieses Links wird in Abb. 33 dargestellt.

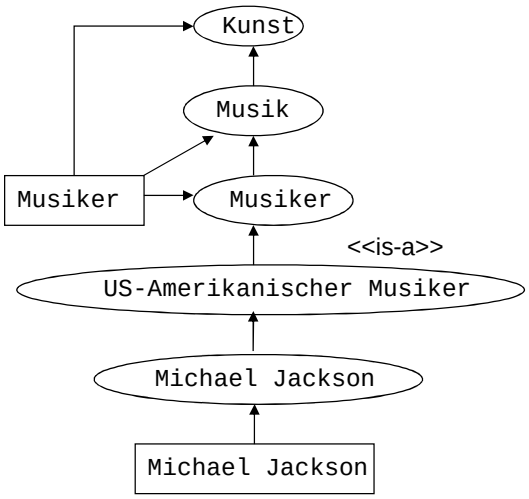

Abbildung 33: Beispiel-Kategoriengraph nach Matching von lexikalischen Köpfen an der richtigen Stelle

Diese Heuristik funktioniert grundsätzlich auch für andere Sprachen. Die Position des lexikalischen Kopfes muss angepasst werden: In Arabisch steht der lexikalische Kopf am Beginn eines Lemmas einer Kategorie. Bei Sprachen mit zusammengesetzten Wörtern wie der deutschen Sprache muss beachtet werden, dass der lexikalische Kopf sich innerhalb dieses zusammengesetzten Wortes befinden kann. Beispielsweise ist das Wort „Baumhaus“ ein solches zusammengesetztes Wort. Diese Heuristik wird für die deutsche Sprache mit Hilfe eines Matching-Fensters der Länge c_1 simuliert. Mit anderen Worten: Ein Kategorienpaar $(c_{i,1},c_{j,1})$ wird als Hyponymie markiert, wenn die letzten c_1-Zeichen der Lemmata identisch sind. Der optimale Wert kann je nach Sprache variieren. Die Experimente für die Parametrisierung auf Deutsch (vgl. Anhang A.1.1) zeigten, dass $c = 4$ die besten Ergebnisse produzierte.

Heuristik 2 (Matching von lexikalischen Köpfen an falscher Stelle)

Die zweite Heuristik basiert auf dem gleichen Ansatz. Lexikalische Köpfe werden aber hier für die Erkennung von Nicht-Hyponymien benutzt. Diese Heuristik überprüft, ob der lexikalische Kopf sich an einer anderen Stelle der Kategorie befindet. Beispielsweise lässt sich mit Hilfe dieser Heuristik erkennen, dass beim Kategorienpaar „Baumhaus“ und „Baum“ keine Hyponymie vorliegt, weil „Baum“ nicht am Ende des Wortes, sondern am Anfang auftritt. Der Algorithmus im Pseudocode wird in Algorithmus 5.1.5) dargestellt.

Algorithmus 5.1.5 Matching von lexikalischen Köpfen an falscher Stelle in Pseudocode

---

Eingabe: Eine Menge von ungelabelten Links $L = \{(c_{1,1},c_{1,2},\varnothing), (c_{2,1},c_{2,2},\varnothing), \ldots, (c_{n,1},c_{n,2},\varnothing)\}$, eine Liste $P = \{p_1), p_2, \ldots, p_k\}$ von Präpositionen in einer gegebenen Sprache

1: **Prozedur** MODIFIER_MATCHING(L,P)
2: **for all** $(c_{i,1},c_{j,2},\varnothing) \in L$ **do**
3: **if** $lemma(c_{i,1})$ enthält $p_a \in P$ **then**
4: $c_{i,1} =$ TRUNCATE_PREPOSITIONS$(c_{i,1},p_a)$
5: **if** $lemma(c_{j,1})$ enthält $p_a \in P$ **then**
6: $c_{j,1} =$ TRUNCATE_PREPOSITIONS$(c_{j,1},p_a)$
7: **if** $lemma(c_{i,1})$ enthält Klammern **then**
8: $c_{i,1} =$ TRUNCATE_BRACKETTS$(c_{i,1})$
9: **if** $lemma(c_{j,1})$ enthält Klammern **then**
10: $c_{j,1} =$ TRUNCATE_BRACKETTS$(c_{j,1})$
11: **if** $lemma(c_{i,1})$ enthält $lemma(c_{j,1})$ nicht an der Position des lexikalischen Kopfes **then**
▷ für Deutsch wird getestet, ob $lemma(c_{i,1}) = lemma(c_{j,1}) \circ w$ gilt. $\circ$ stellt die Konkatenation von zwei Zeichen dar und w eine beliebige Zeichenkette
12: $(c_{i,1},c_{j,2},\varnothing) \leftarrow (c_{i,1},c_{j,2},0)$ ▷ Link wird als Nicht-Hyponymie markiert

Ausgabe: Menge von gelabelten Links $L' = \{(c_{1,1},c_{1,2},0), (c_{2,1},c_{2,2},0), \ldots, (c_{n,1},c_{n,2},0)\}$

Auch diese Heuristik eignet sich für verschiedene Sprachen, wenn man ähnlich wie oben beschrieben ein weiteres Matching-Fenster c_2 definiert. Für Deutsch hat beispielsweise $c_2 = 6$ die besten Ergebnisse geliefert (vgl. Anhang A.1.1). Für andere Sprachen lassen sich einfach Wörter oder Teilwörter benutzen. Im exemplarischen Beispiel (siehe Abb. 34) führt diese Heuristik dazu, dass der Link (`Musik`,`Musiker`) als Nicht-Hyponymie erkannt wird.

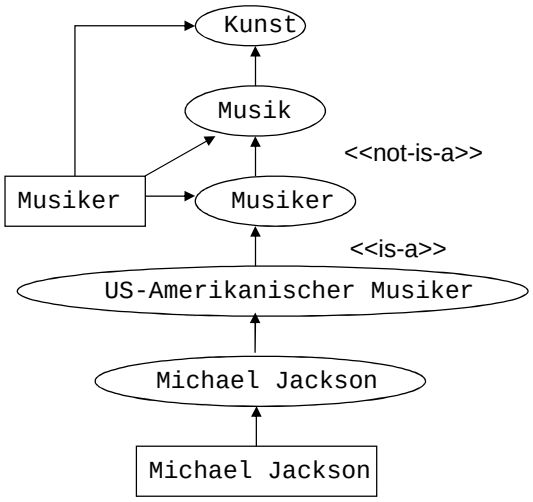

Abbildung 34: Beispiel-Kategoriengraph nach Matching von lexikalischen Köpfen an anderen Stellen

Heuristik 3 (Matching des ersten Satzes eines Artikels)

Die dritte Heuristik basiert auf der Tatsache, dass die Glosse eines Wikipedia-Artikels (vgl. Abschnitt 2.4.2.1) eine besondere Bedeutung für Hyponymie-Beziehungen hat. Anstatt sprachabhängige Patterns (vgl. Abschnitt 3.2.2.2) für jede Sprache zu definieren, versucht diese Heuristik im ersten Satzes des Artikels [109] eine Hyponymie-Beziehung zwischen Kategorien zu finden.

Im exemplarischen Kategoriengraph lässt sich mittels der Heuristik bestimmen, dass (`Michael Jackson`,`US-amerikanischer Musiker`,1) gilt, da der erste Satz des Wikipedia-Artikels „Michael Jackson“ wie folgt lautet:

„Michael Joseph Jackson (geboren am 29. August 1958 in Gary, Indiana; Gestorben am 25. Juni 2009 in Los Angeles, Kalifornien) war ein US-amerikanischer Musiker, Komponist, Tänzer und Entertainer.“[1]

Algorithmus 5.1.6 Matching des ersten Satzes eines Artikels

Eingabe: Eine Menge von ungelabelten Links $L = \{(c_{1,1},c_{1,2},\varnothing), (c_{2,1},c_{2,2},\varnothing), \ldots, (c_{n,1},c_{n,2},\varnothing)\}$, die einen gleichnamigen Wikipedia-Artikel $a(lemma(c_{i,1}))$ haben

1: **Prozedur** FIRST_SENTENCE_MATCHING(L,$a(lemma(c_{i,1}))$)
2: **for all** $(c_{i,1},c_{j,2},\varnothing) \in L$ **do**
3: first_sentence = GET_FIRST_SENTENCE($a(c_{i,1})$)
4: **if** first_sentence enthält Teilstring $lemma(c_{j,2})$ **then**
5: $(c_{i,1},c_{j,2},\varnothing) \leftarrow (c_{i,1},c_{j,2},1)$ ▷ Link wird als Hyponymie markiert

Ausgabe: Menge von gelabelten Links $L' = \{(c_{1,1},c_{1,2},1), (c_{2,1},c_{2,2},1), \ldots, (c_{n,1},c_{n,2},1)\}$

Der große Vorteil dieser Heuristik ist die Tatsache, dass eine sprachabhängige Suche nach Pattern nicht benötigt wird, da die Wikipedia-Guidelines in allen Sprachen definieren, dass der erste Satz den Artikel definieren soll.

1 `http://de.wikipedia.org/wiki/Michael_Jackson` - Zugriff am 14.11.2012

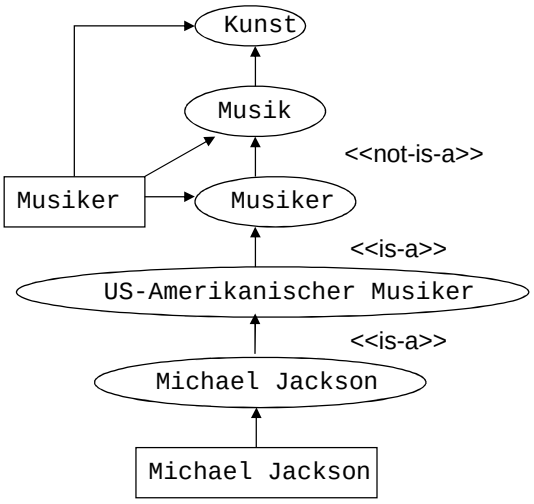

Abbildung 35: Beispiel-Kategoriengraph nach Matching des ersten Satzes

Heuristik 4 (Matching von Kookkurrenzen im Graph)

Diese vierte Heuristik basiert auf der Ausnutzung der Struktur des Kategoriengraphen. Ponzetto et al. [118] wiesen darauf hin, dass Kookkurrenzen im Graph auf *is-a*-Relationen zwischen Kategorien hindeuten können. Dies gilt insbesondere, wenn ein Artikel sowohl in einer Kategorie als auch in einer Unterkategorie enthalten ist. Im Beispiel führt diese Heuristik dazu, dass (`Musik`,`Kunst`,1) gilt, da beide Kategorien den gemeinsamen Artikel „Musiker“ haben.

Algorithmus 5.1.7 Matching von Kookkurrenzen im Graph

Eingabe: Eine Menge von ungelabelten Links $L = \{(c_{1,1},c_{1,2},\varnothing), (c_{2,1},c_{2,2},\varnothing), \ldots, (c_{n,1},c_{n,2},\varnothing)\}$, $A(c_{i,j})$ für $c_{i,j} \in L$

1: **Prozedur** CO-OCCURRENCES_MATCHING(L,)
2: **for all** $(c_{i,1},c_{j,2}) \in L$ **do**
3: **if** $A(c_{i,1}) \cap A(c_{j,2}) \neq \emptyset$ **then**
4: $(c_{i,1},c_{j,2},\varnothing) \leftarrow (c_{i,1},c_{j,2},1)$ ▷ Link wird als Hyponymie markiert

Ausgabe: Menge von gelabelten Links $L' = \{(c_{1,1},c_{1,2},1), (c_{2,1},c_{2,2},1), \ldots, (c_{n,1},c_{n,2},1)\}$

Diese Heuristik kann unabhängig von der Sprache angewendet werden. Das Ergebnis dieser Heuristik wird in der folgenden Abbildung dargestellt.

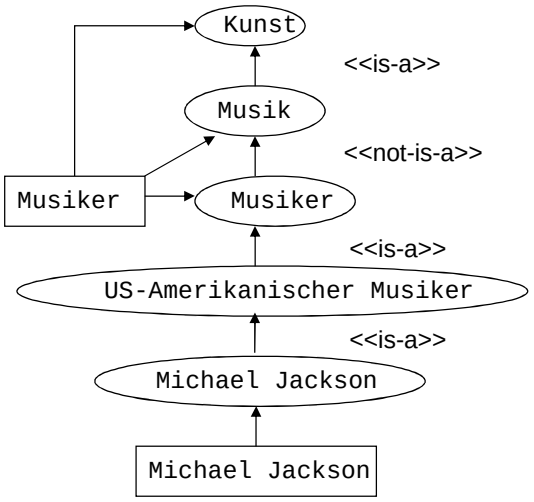

Abbildung 36: Beispiel-Kategoriengraph nach Matching von Kookkurrenzen

5.1.2.3 *Nachverarbeitungsschritte*

Nach Anwendung der oben erwähnten Heuristiken müssen gefundene Relationen transitiv weiter übertragen werden. Im Beispiel-Kategoriengraph wird aufgrund der Übersichtlichkeit in der graphischen Darstellung auf *not-is-a*-Relationen und auf Artikel verzichtet. Es bleiben zwei Untergraphen, die in Abb. 37 dargestellt werden.

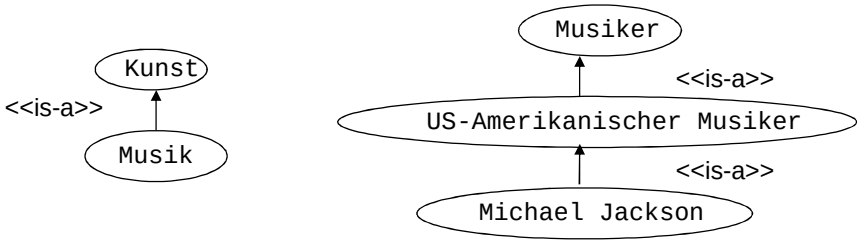

Abbildung 37: Beispiel-Kategoriengraph ohne *not-is-a*-Relationen und Artikel

Nachrverarbeitungsschritt 1 (Propagierung der gefundenen Links)

Dieser Nachverarbeitungsschritt markiert ein Kategorienpaar (c_1,c_2) als *is-a*, wenn eine Kategorie c_l existiert, so dass (c_1,c_l) und (c_l,c_2) gilt. Es entspricht der mathematischen Relation der Transitivität.

Algorithmus 5.1.8 Propagierung von gefundenen Links

Eingabe: Eine Menge von Links $L = \{(c_{1,1},c_{1,2},1), (c_{2,1},c_{2,2},1), \ldots , (c_{n,1},c_{n,2},1)\}$, die Liste der Kategorien $C = \{c_1,c_2,\ldots,c_l\}$

1: **Prozedur** LINK_PROPAGATION_RULE(L,)
2: **for all** $(c_{a,1},c_l) \in L$ **do**
3: **for all** $(c_l,c_{b,2}) \in L$ **do**
4: CREATE_IS-A-RELATION$(c_{i,1},c_{j,2})$ ▷ Link wird erstellt und als Hyponymie markiert

Ausgabe: Menge von gelabelten Links $L' = \{(c_{1,1},c_{1,2},1), (c_{2,1},c_{2,2},1), \ldots ,(c_{n,1},c_{n,2},1)\}$

Im Beispiel ergänzt dieser Schritt die Relation (`Michael Jackson`,`Musiker`,1). Diese Heuristik baut auf bereits vorhandenen Relationen auf und ist somit sprachunabhängig. Dieser Schritt ist der einzige Schritt, der neue (noch nicht) bekannte Links erzeugt.

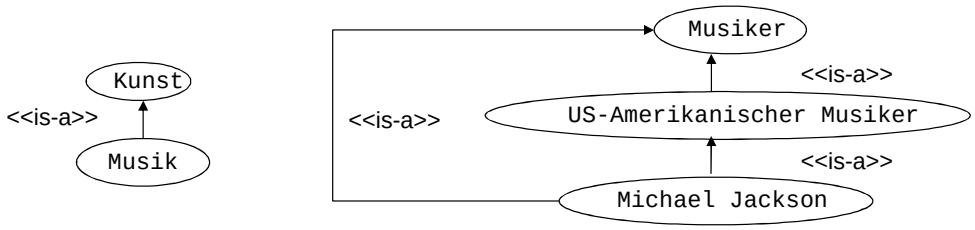

Abbildung 38: Beispiel-Kategoriengraph nach Propagierung von Hyponymie-Beziehungen

Die Ausgabe des Verfahrens sieht am Ende wie folgt aus:
$Q = \{$((`Kunst`,`Musik`,1), (`Musik`,`Musiker`,0), (`Musiker`,`US-Amerikanischer Musiker`,1), (`Michael Jackson`,`Musiker`,1),(`Michael Jackson`,`US-Amerikanischer Musiker`,1)$\}$

5.1.3 *Sprachunabhängigkeit des Verfahrens*

Der hier vorgestellte Ansatz (TaxWikiHeur.KOM) ist insofern sprachunabhängig, da er in verschiedenen Sprachen ohne große Modifizierung der Heuristiken angewendet werden kann. Der Ansatz kann benutzt werden, um eine große Menge an Hyponymierelationen zu erkennen, wozu nur geringe sprachspezifische Modifikationen notwendig sind.

Die benötigten Modifikationen ergeben sich aus den Eingaben, die die verschiedenen Schritte brauchen. Wenn das Verfahren auf eine Sprache S angewendet werden soll, braucht TaxWikiHeur.KOM folgende Angaben:

1. Eine Liste von Präfixen, die Wikipedia in der Sprache S benutzt, um Metakategorien zu erkennen, wie z.B. „Wikipedia“ oder „Benutzer“(siehe Abschnitt 5.1.2.1).

2. Die Präposition, die in S benutzt wird, um Verfeinerungslinks zu definieren. Beispielsweise „by“ für Englisch oder „nach“ für Deutsch (siehe Abschnitt 5.1.2.1).

3. Die Liste aller Präpositionen in Sprache S, mittels derer lexikalische Köpfe in komplexen Kategorien heuristisch bestimmt werden können (siehe Abschnitt 5.1.2.2).

4. Die Information darüber, wo der lexikalische Kopf in S zu finden ist (siehe Abschnitt 5.1.2.2).

Die Nutzbarkeit des Verfahrens in verschiedenen Sprachen wird im nächsten Abschnitt nachgewiesen.

5.1.4 *Evaluation des Verfahrens*

TaxWikiHeur.KOM wurde in vier verschiedene Sprachen evaluiert: Drei europäische Sprachen (Englisch, Deutsch und Spanisch) und eine Sprache mit nicht-lateinischer Schrift (Arabisch). Die Evaluation bestand aus drei Schritten:

1. Erstellung von manuell gelabelten Korpora für die vier verschiedenen Sprachen (Abschnitt 5.1.4.1). Hierbei wird aus dem Kategoriengraph der Wikipedia zufällig eine kleine Teilmenge extrahiert. Jeder Link aus dem Korpus wird manuell gelabelt, d.h. jedes Paar wird annotiert, je nachdem, ob es sich um eine Hyponymierelationen handelt oder nicht. Dieser Korpus wird sowohl für die Überprüfung im nächsten Schritt als auch zur Parametrisierung einiger Heuristiken verwendet.

2. Überprüfung der durch die Heuristiken erstellten Klassifikation mit Hilfe der manuell gelabelten Korpora (Abschnitt 5.1.4.2). Die Links-Korpora aus Schritt (1) werden benutzt, um die Güte der Klassifikation für jede Sprache zu bestimmen.

3. Vergleich zu externen manuell erstellten Wissensbasen (WordNet, GermaNet) bei der Anwendung der Heuristiken auf dem gesamten Kategoriengraph (Abschnittsec:eval-ext-kb). Dieser Schritt soll Aufschluss darüber geben, wie gut die Abdeckung des Verfahrens ist.

Im Folgenden sollen die Ergebnisse jedes Schrittes vorgestellt werden.

5.1.4.1 *Erstellung von manuell gelabelten Korpora für die vier verschiedenen Sprachen*

Der Korpus besteht aus Links einer zufällig gewählten Menge von 1000 Wikipedia-Artikeln. Der Korpus wird unter Benutzung der „zufälliger Artikel"-Funktion[2], der Wikipedia extrahiert. Diese Funktion steht nicht nur für Deutsch, sondern für alle anderen Sprachen zur Verfügung.

Für jeden zufällig gewählten Artikel werden die Kategorien und deren Oberkategorien als Paar extrahiert. Diese Paare werden mit Hilfe der Wikipedia-Exportierungsseiten[3] als XML-Datei extrahiert. Diese Funktion existiert ebenfalls für alle anderen Sprachen. Abschließend werden die Links in eine CSV[4]-Datei[5] gespeichert und manuell annotiert (mit der Information, ob es sich um Hyponymierelationen handelt oder nicht). Ein Überblick der extrahierten Korpora wird in Tabelle 2 dargestellt.

2 `http://de.wikipedia.org/wiki/Special:Random` - Zugriff am 14.11.2012
3 `http://de.wikipedia.org/wiki/Special:Export` - Zugriff am 14.11.2012
4 Comma-separated values (CSV)
5 `http://tools.ietf.org/html/rfc4180` - Zugriff am 14.11.2012

Tabelle 2: Überblick über die manuell extrahierten und gelabelten Korpora

Language	Englisch	Spanisch	Deutsch	Arabisch
Anzahl *is-a*-Links	454 (29,1 %)	331 (38,5 %)	157 (18,7 %)	301 (25,0 %)
Anzahl *not-is-a*-Links	1107 (70,9 %)	529 (61,5 %)	575 (81,3 %)	903 (75,0 %)
Gesamtanzahl Links	1561	860	732	1204

5.1.4.2 *Überprüfung der durch die Heuristiken erstellten Klassifikation mit Hilfe der manuell gelabelten Korpora*

Die zuvor vorgestellten Heuristiken werden nacheinander auf die verschiedenen Korpora angewendet. In diesem Kapitel werden die Ergebnisse für die deutsche Sprache beschrieben. Die Ergebnisse der anderen Sprachen sind in Anhang A.1.2 dargestellt.

Filtern von administrativen Metakategorien: Diese Heuristik wird auf die extrahierten Korpora angewendet. Die Anwendung führt dazu, dass Links, in denen sich administrative Präfixe befanden, als *not-is-a* markiert wurden. Für die deutsche Sprache blieben nach diesem Vorverarbeitungsschritt 726 Links ungelabelt.

Filtern von Verfeinerungslinks: Durch die Regeln zur Namensgebung in Wikipedia kann diese Heuristik in allen Sprachen ohne Fehler durchgeführt werden. Diese Heuristik markiert die betreffenden Links als *not-is-a*. Nach Anwendung dieses Vorverarbeitungsschritts blieben 495 Links, die noch nicht klassifiziert wurden.

Matching von lexikalischen Köpfen an richtiger Stelle: Bei der Anwendung dieser Heuristik wurden 121 Links als *is-a*-Links markiert. Von diesen 121 Links wurden 120 korrekt und 1 falsch klassifiziert. Dies bestätigt die Arbeit in [118], wo diese Heuristik auch als sehr gut abgeschnitten hat.

Matching von lexikalischen Köpfen an falscher Stelle: Diese Heuristik hat in allen Sprachen eine gute Performance gezeigt. Bei der Anwendung dieser Heuristik auf Deutsch wurden 20 Links als *not-is-a*-Links markiert. Von diesen 20 Links wurden 17 korrekt und 3 falsch klassifiziert. Es bleiben weitere 354 noch nicht markierte Links.

Matching des ersten Satzes eines Artikels: Diese Heuristik konnte überraschenderweise wenige Links in den betrachteten Sprachen finden. Die gefundenen Links wurden allerdings in den meisten Fällen (8 von 9) richtig klassifiziert.

Matching von Kookkurrenzen im Graph: Diese Heuristik markierte 24 Links als *is-a*, wobei 18 davon korrekt klassifiziert wurden. Vor dem letzten Schritt bleiben 322 ungelabelte Links. Diese Links konnte also keine der Heuristiken labeln.

Propagierung der gefundenen Links: Zum Schluss wurden neue Links propagiert, sodass 141 neue Links hinzugekommen sind. Bei der manuellen Überprüfung dieser Links wurde festgestellt, dass es sich 110 neue korrekte Links gab und 31 nicht korrekt waren. Die verbleibenden ungelabelten Links wurden als *not-is-a*-Links markiert.

Tabelle 3: Klassifikationsergebnisse der einzelnen Heuristiken für Deutsch

	Ungelabelte Links	Korrekt klassifiziert	Inkorrekt klassifiziert	Noch zu klassifizieren
Heuristik: Admin. Metakategorien	732	6	0	726
Heuristik: Verfeinerungslinks	726	231	0	495
Heuristik: Lex. Köpfen an richtiger Stelle	495	120	1	374
Heuristik: Lex. Köpfen an falscher Stelle	374	17	3	354
Heuristik: Erster Satz eines Artikels	354	8	1	345
Heuristik: Kookkurrenzen im Graph	345	18	6	322
Heuristik: Transitive Links	322	110	31	322

Zusammengefasst ergaben sich folgende Ergebnisse:

Tabelle 4: Zusammenfassung der Ergebnisse von TaxWikiHeur.KOM

Sprache	Englisch	Spanisch	Deutsch	Arabisch
Korrekt klass. Links	1205 (72,6 %)	635 (68,8 %)	711 (90,5 %)	1022 (83,3 %)
Falsch klass. Links	455 (27,4 %)	289 (31,2 %)	74 (9,5 %)	205 (16,7 %)
Gesamtanzahl	1660	924	785	1227

In Tabelle 5 wird ein näherer Blick auf die Ergebnisse geworfen. Für die deutsche Sprache lag die Precision bei 93,6 % und Recall bei 84,2 % für *is-a*-Relationen. Während die Precision in allen Sprachen relativ gut war, war dies für den Recall nicht der Fall. Dies weist darauf hin, dass das Verfahren gute Ergebnisse für klassifizierte Links produziert, aber darunter leidet, dass viele Links durch die Heuristiken nicht erfasst werden.

Tabelle 5: Precision, Recall und F_1-Maß für jede Klasse und Sprache

	Precision	Recall	F_1-Maß	Klasse
Englisch	53,2 %	34,1 %	41,5 %	*is-a*
	76,9 %	88,0 %	82,0 %	*not-is-a*
Spanisch	67,5 %	37,1 %	47,9 %	*is-a*
	69,0 %	88,6 %	77,6 %	*not-is-a*
Deutsch	84,1 %	78,3 %	81,1 %	*is-a*
	92,6 %	94,8 %	93,7 %	*not-is-a*
Arabisch	76,1 %	48,0 %	58,9 %	*is-a*
	84,6 %	95,0 %	89,5 %	*not-is-a*

Im nächsten Schritt soll TaxWikiHeur.KOM auf alle Links im Kategoriengraph der Wikipedia angewendet werden, um eine genauere Beurteilung der Performanz des Verfahrens zu bekommen. Dies geschieht im nächsten Abschnitt durch den Vergleich mit externen manuell erstellten Wissensbasen.

5.1.4.3 *Vergleich zu externen manuell erstellten Wissensbasen*

Zur Beurteilung der Taxonomien erfolgte ein Vergleich der Klassifikation aller Wikipedia-Links mit manuell erstellten Wissensbasen. Die Genauigkeit von TaxWikiHeur.KOM wird mit Hilfe von WordNet und GermaNet verglichen. Zu diesem Zweck wurde die Wikipedia-Version vom 30.11.2010 genutzt. Der Kategoriengraph auf der deutschen Wikipedia besteht aus 80198 Kategorien und 168353 Links und auf der englischen Wikipedia aus 1111805 Kategorien und 1.085.254 Links.

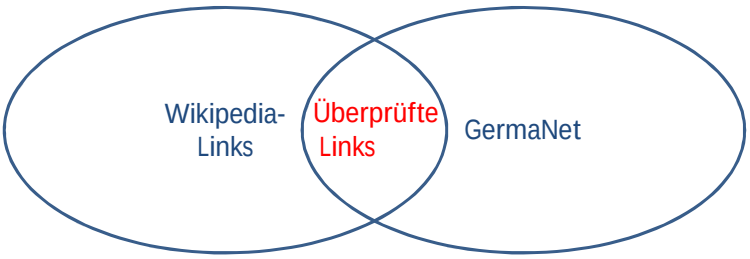

Abbildung 39: Links, die sowohl im Kategorien-Graph als auch in GermaNet vorkommen

Tabelle 6 zeigt einen Überblick der Ergebnisse. Sie zeigt korrekt und inkorrekt klassifizierte Links. Für die Deutsche Wikipedia wurden 90,2 % der gelabelten Links korrekt und 9,8 % falsch klassifiziert. Allerdings konnte nur der Teil der Wikipedia evaluiert werden, der eine Überlappung mit GermaNet (bzw. WordNet für Englisch) hat (siehe Abb. 39). Ein Link (c_1,c_2)wurde genau dann evaluiert, wenn sowohl c_1 als auch c_2 als Kategorie (Wikipedia) und Konzept (GermaNet und WordNet) existieren. Aus diesem Grund konnten 80.551 (42,8 %) der Instanzen nicht überprüft werden. Links, die in einer der Wissensbasen nicht oder nur teilweise vorkommen, wurden als „unbekannte Links“ markiert.

Tabelle 6: Vergleich der Ergebnisse mit GermaNet und WordNet

	Deutsch	Englisch
Korrekt klassifizierte Links	72731 (90,2 %)	439280 (86,8 %)
Falsch klassifizierte Links	7820 (9,8 %)	58550 (13,2 %)
Unbekannte Links	87802 (52,2 %)	613975 (58,1 %)
Gesamtanzahl Links	168353	1111805

In Tabelle 7 wird ein näherer Blick auf die Precision, Recall und F_1-Maß der Links geworfen, die mit WordNet und GermaNet evaluiert werden konnten. Für die deutsche Sprache lag bei *is-a*-Relationen die Precision bei 93,6 % und Recall bei 84,2 %.

Tabelle 7: Precision, Recall und F_1-Maß beim Vergleich der Ergebnisse mit GermaNet und WordNet

	Precision	Recall	F_1-Maß	Klasse
Englisch	52,9 %	88,2 %	73,4 %	is-a
	98,9 %	40,7 %	69,8 %	not-is-a
Deutsch	93,6 %	84,2 %	88,6 %	is-a
	87,9 %	95,2 %	91,4 %	not-is-a

Bei der Analyse der Ergebnisse muss man beachten, dass obwohl die Ergebnisse sehr gut sind, sie sich nur auf den Teil der Wikipedia-Links beziehen, die mit GermaNet evaluiert werden konnten. Insgesamt wurden 80.551 Links in der deutschen Sprache überprüft, aber 87.802 Link blieben unüberprüft, da keine vollständige In-

formationen über den Link in GermaNet zu finden waren. Das hat zwei Ursachen: Erstens hat Wikipedia eine viel größere Abdeckung als manuell erstellte Wissensbasen und zweitens konnten einige Begriffe nicht gefunden werden, weil sie in einer anderen Schreibweise in GermaNet/WordNet vorliegen. Beispielsweise finden sich die Plural-Formen von manchen Begriffen aus der Wikipedia in GermaNet/WordNet nicht, weil sie nur in der Singular-Form vorliegen.

Schließlich überrascht die Tatsache, dass die Precision für Deutsch und Englisch so unterschiedlich ist. Dies war auch beim Vergleich mit dem manuell erstellten Korpus beobachtet worden. Hier fiel auf, dass die parametrisierten Heuristiken verhältnismäßig mehr Links in Deutsch als im Englisch erfassen konnten und die erfassten Links wurden mit einer hohen Qualität klassifiziert. Für die anderen Sprachen gab es keine Möglichkeit (abgesehen von der Position des lexikalischen Kopfes) der Parametrisierung einzelner Heuristiken. Weitere Anpassungen von TaxWikiHeur.KOM an eine bestimmte Sprache können dazu führen, dass die Ergebnisse für diese Sprache verbessert werden.

5.1.5 *Zusammenfassung*

In diesem Abschnitt wurde ein multilinguales Verfahren zur Klassifikation von Links im Wikipedia-Kategoriengraph in *is-a-* und *not-is-a-*Relationen vorgestellt und evaluiert. Es wurde gezeigt, dass Hyponymie-Beziehungen zwischen Wikipedia-Kategorien automatisch mit wenigen Informationen über eine Sprache erkannt werden können. Die Ergebnisse des Verfahren wurden mit Hilfe der manuell erstellen Korpora überprüft. Als Erstes wurde zu diesem Zweck ein zufällig erstellter manuell gelabelter Korpus verwendet und anschließend die von Experten erstellten Wissensbasen GermaNet und WordNet. Die Ergebnisse auf dem manuell erstellten Korpus sind relativ gut, allerdings war der Recall immer kleiner als die Precision, was sich dadurch erklären lässt, dass viele Links nicht durch eine Heuristik erfasst werden konnten.

5.2 ERKENNUNG VON HYPONYMIEN AUF BASIS VON ENTSCHEIDUNGSBÄUMEN

Im vorigen Abschnitt wurde gezeigt, dass die automatische multilinguale Erkennung von Hyponymierelationen im Kategoriengraph grundsätzlich möglich ist. Basierend auf den im vorigen Kapitel gewonnenen Erkenntnissen, wird in den folgenden Abschnitten eine robustere Methode vorgestellt. Zum einen soll das Verfahren auf maschinellem Lernen beruhen und auf diese Weise sicherstellen, dass alle Links durch jedes Feature erfasst werden. Zum anderen werden die Klassifikatoren mit verschiedenen Korpora trainiert, so dass der Klassifikator für jede Sprache anders aussehen wird. Somit erreicht man eine Art „Parametrisierung“ des Verfahrens. Das in diesem Abschnitt vorgestellte Verfahren trägt den Namen TaxWikiML.KOM und verwendet 20 Features, um Hyponymierelationen im Kategoriengraph der Wikipedia zu erkennen [37].

5.2.1 *Features*

Die Eingabe von TaxWikiML.KOM ist, genau wie bei TaxWikiHeur.KOM, eine Menge von Links $l = (c_1, c_2)$ aus dem Kategoriengraph der Wikipedia. Diese Links werden

in die zwei Klassen *is-a* oder *not-is-a* klassifiziert (siehe Abschnitt 2.2.2). Für jeden Link wird ein Featurevektor aus Werten der 20 Features gebildet. In Tabelle 8 sind die entwickelten Features im Überblick dargestellt. Nachfolgend werden die einzelnen Features im Detail erklärt.

Tabelle 8: Überblick der entwickelten Features

ID	Name	Wertebereich	Feature-Typ
1	`adminCatFeature`	{0,1}	Vorverarbeitung
2	`refinementLinkFeature`	{0,1}	Vorverarbeitung
3	`positionOfHeadFeature`	{2,1,0,-1}	Syntaktisch
4	`coocurrenceOfWordsFeature`	$\mathbb{N}$	Syntaktisch
5	`coocurrenceArticleFeature`	{0,1}	Strukturell
6	`commonArticleFeature`	{1,0,-1}	Strukturell
7	`c1c2IncomingLinksFeature`	{1,0,-1}	Strukturell
8	`c1c2OutgoingLinksLinksFeature`	{1,0,-1}	Strukturell
9	`c1distanceCommonAncestorFeature`	$\mathbb{N}$	Strukturell
10	`c2distanceToCommonAncestorFeature`	$\mathbb{N}$	Strukturell
11	`c1NumberOfSubcategoriesFeature`	$\mathbb{N}$	Strukturell
12	`c1NumberOfSuperCategoriesFeature`	$\mathbb{N}$	Strukturell
13	`c2NumberOfSubcategoriesFeature`	$\mathbb{N}$	Strukturell
14	`c2NumberOfSuperCategoriesFeature`	$\mathbb{N}$	Strukturell
15	`CommonWikilinksFeature`	$\mathbb{N}$	Strukturell
16	`firstSentenceFeature`	{0,1}	Artikelbasiert
17	`RedirectFeature`	{1,0,-1}	Artikelbasiert
18	`c2Inc1Feature`	$\mathbb{N}$	Artikelbasiert
19	`c1ArticleFeature`	{0,1}	Artikelbasiert
20	`c2ArticleFeature`	{0,1}	Artikelbasiert

5.2.1.1 *Vorverarbeitungsfeatures*

Es existieren zwei Vorverarbeitungsfeatures, die grundsätzlich den Vorverarbeitungsheuristiken in Abschnitt 5.1.2.1 entsprechen. Beide Features liefern den Wert 1 für Links, die administrative Kategorien (`adminCatFeature`) oder Verfeinerungslinks (`refinementLinkFeature`) enthalten, zurück. Die Berechnung der Features erfolgt nach den in Algorithmen 5.2.1 und 5.2.2 beschriebenen Prozeduren. Diese Features können, wie in Abschnitt 5.1.2.1 erklärt, in verschiedenen Sprachen berechnet werden.

Algorithmus 5.2.1 Berechnung von adminCatFeature

Eingabe: Link $l = \{(c_1,c_2)\}$, eine Liste $P = \{p_1, p_2, \ldots, p_k\}$ von Präfixen, die Metakategorien darstellen

```
1: function adminCatFeature(l,P)
2:     if lemma(c1) or lemma(c2) begins with pa ∈ P then
3:         return 1
4:     else
5:         return 0
```

Ausgabe: Ein Wert $w \in \{0,1\}$

Algorithmus 5.2.2 Berechnung von refinementLinkFeature

Eingabe: Link $l = \{(c_1,c_2)\}$, die Präposition p, die in einer gegebenen Sprache zur Angabe von Verfeinerungslinks benutzt wird

```
1: function refinementLinkFeature(L,p)
2:     if lemma(c1) or lemma(c2) enthält String " p " then
3:         return 1
4:     else
5:         return 0
```

Ausgabe: Ein Wert $w \in \{0,1\}$

5.2.1.2 *Syntaktische Features*

Syntaktische Features nutzen die syntaktischen Komponenten der Kategorien, um zwischen *is-a*- und *not-is-a*-Relationen zu unterscheiden (siehe Abschnitt 3.2.2.1). In diesem Abschnitt werden zwei syntaktische Features vorgestellt: `positionOfHeadFeature` und `coocurrenceOfWordsFeature`. Das `positionOfHeadFeature` nutzt die Tatsache, dass der lexikalische Kopf der Lemmata sehr effektiv zur Erkennung von Hyponymien [118] verwendet werden kann. Dieses Feature liefert einen Wert zwischen $\{2,1,0,-1\}$ für einen Link $l = \{(c_1,c_2)\}$ zurück, der die Position des lexikalischen Kopfes von c_2 repräsentiert. Wir unterscheiden folgende Fälle:

$$f_3(c_1,c_2) = \begin{cases} 2 & \text{Wenn der lexikalische Kopf von } c_2 \text{ am Ende von } c_1 \text{ steht} \\ 1 & \text{Wenn der lexikalische Kopf von } c_2 \text{ irgendwo in der Mitte von } c_1 \text{ steht} \\ 0 & \text{Wenn der lexikalische Kopf von } c_2 \text{ am Anfang von } c_1 \text{ steht} \\ -1 & \text{sonst, d.h. der lexikalische Kopf kommt gar nicht vor} \end{cases}$$

Dieses Feature wird mit Hilfe von Algorithmus 5.2.3 berechnet:

Beispielsweise berechnet dieser Algorithmus für den Link $l = \{(\text{Französische Revolution},\text{Revolution})\}$ den Wert 2. In Abschnitt 5.1.2.2 wurde angegeben, dass Kategorien Präpositionen oder Klammern enthalten können. Diese Fälle werden entsprechend gesondert behandelt. Es wird, genau wie bei TaxWikiHeur.KOM, angenommen, dass, wenn die Position des lexikalischen Kopfes nicht mit der Position des lexikalischen Kopfes in einer gegebenen Sprache übereinstimmt, es sich bei diesem Link um eine *not-is-a*-Relation handelt.

Algorithmus 5.2.3 Berechnung von positionOfHeadFeature

Eingabe: Link $l = \{(c_1,c_2)\}$, eine Liste $P = \{p_1), p_2, \ldots, p_k\}$ von Präpositionen in einer gegebenen Sprache

function POSITIONOFHEADFEATURE(L,P)
  if $lemma(c_1)$ enthält $p_a \in P$ **then**
    $c_1 =$ TRUNCATE_PRESPOSITIONS(c_1,p_a)
  if $lemma(c_2)$ enthält $p_a \in P$ **then**
    $c_2 =$ TRUNCATE_PRESPOSITIONS(c_2,p_a)
  if $lemma(c_1)$ enthält Klammern **then**
    $c_1 =$ TRUNCATE_BRACKETTS(c_1)
  if $lemma(c_2)'$ enthält Klammern **then**
    $c_2 =$ TRUNCATE_BRACKETTS(c_2)

▷ Die Funktionen Truncate_Prepositions und Truncate_Bracketts sind wie in Abschnitt 5.1.2.2 beschrieben

  if $lemma(c_1)$ endet mit $lemma(c_2)$ **then**
    return 2
  else if $lemma(c_1)$ beginnt mit $lemma(c_2)$ **then**
    return 0
  else if $lemma(c_1)$ enthält $lemma(c_2)$ **then**
    return 1
  else
    return -1

▷ z.B. $lemma(c_1) = w \circ lemma(c_2)$ auf Deutsch/Englisch oder $lemma(c_1) = lemma(c_2) \circ w$ auf Spanisch/Arabisch, wobei w eine beliebige Zeichenkette darstellt

Ausgabe: Ein Wert $w \in \{2,1,0,-1\}$

`cooccurrenceOfWords` bestimmt Kookkurrenzen von Wörtern in beiden Kategorienamen. Dieses Feature soll Fälle behandeln, in denen die Lemmata zweier verlinkter Kategorien mehrere Wörter gemeinsam haben. Die Funktion zur Berechnung von `cooccurrenceOfWords` wird in Algorithmus 5.2.4 dargestellt.

Algorithmus 5.2.4 Berechnung von cooccurrenceOfWords

Eingabe: Link $l = \{(c_1,c_2)\}$, eine Liste $P = \{p_1, p_2, \ldots, p_k\}$ von Präpositionen in einer gegebenen Sprache

function COOCCURRENCEOFWORDSFEATURE(L,P)
  if $lemma(c_1)$ enthält $p_a \in P$ **then**
    $c_1 =$ TRUNCATE_PREPOSITIONS(c_1,p_a)
  if $lemma(c_2)$ enthält $p_a \in P$ **then**
    $c_2 =$ TRUNCATE_PREPOSITIONS(c_2,p_a)
  if $lemma(c_1')$ enthält Klammern **then**
    $c_1 =$ TRUNCATE_BRACKETTS(c_1)
  if $lemma(c_2)'$ enthält Klammern **then**
    $c_2 =$ TRUNCATE_BRACKETTS(c_2)
  $occ = 0$
  for all $w \in lemma(c_1)$ **do**
    if $w \in lemma(c_2)$ **then**
      $occ = occ + 1$
  if $occ \geqslant 2$ **then**
    return 1
  else
    return 0

Ausgabe: Ein Wert $w \in \{0,1\}$

5.2.1.3 *Strukturelle Features*

Diese Features nutzen die Struktur des Kategoriengraphs zusammen mit dem Wikilinkgraph[6]. `coocurrenceFeature` liefert den Wert 1 für Links, dessen Kategorien mindestens einen Artikel gemeinsam haben, ansonsten liefert das Feature den Wert 0 zurück. Die Berechnung von `coocurrenceFeature` wird im Algorithmus 5.2.5 dargestellt.

6 Die Struktur der Wikipedia lässt sich als Graph darstellen, wenn Artikel als Knoten und Wikilinks als Kanten angesehen werden.

Algorithmus 5.2.5 Berechnung von coocurrenceFeature

Eingabe: Link $l = \{(c_1, c_2)\}$

```
1: function COOCURRENCEFEATURE(L)
2:     A_1 = a(c_1)                    ▷ a(c) berechnet alle Artikel zu einer Kategorie.
3:     A_2 = a(c_2)
4:     if A_1 ∩ A_2 ≠ ∅ then
5:         return 1
6:     else
7:         return 0
```

Ausgabe: Eine natürliche Zahl

Darüber hinaus wird die Anzahl der gemeinsamen Artikel durch ein eigenes Feature (*commonArticleFeature*) berechnet (vgl. Algorithmus 5.2.6).

Algorithmus 5.2.6 Berechnung von commonArticleFeature

Eingabe: Link $l = \{(c_1, c_2)\}$

```
1: function COMMONARTICLEFEATURE(L)
2:     A_1 = a(c_1)                    ▷ a(c) berechnet alle Artikel zu einer Kategorie.
3:     A_2 = a(c_2)
4:     occ = |{A_1 ∩ A_2}|
5:     return occ
```

Ausgabe: Eine natürliche Zahl

`c1c2IncomingLinksFeature` und `c1c2OutgoingLinksFeature` messen die Stärke der Relation zwischen beiden Kategorien. Zu diesem Zweck wird die Anzahl der Artikel in c_1 gezählt, die mindestens einen ein- oder ausgehenden Wikilink zu einem Artikel in c_2 haben [30]. Die Funktionen zur Berechnung dieser Features werden in Algorithmen 5.2.7 und 5.2.8 vorgestellt.

Algorithmus 5.2.7 Berechnung von ausgehenden Links

Eingabe: Link $l = \{(c_1, c_2)\}$

```
1: function C1C2OUTCOMINGLINKSFEATURE(L)
2:     occ = 0
3:     for all a_i ∈ c_1 do
4:         if a_i hat ein Wikilink zu einem Artikel a_j ∈ c_2 then
5:             occ = occ + 1
6:     return occ
```

Ausgabe: Eine natürliche Zahl

Algorithmus 5.2.8 Berechnung von eingehenden Links

Eingabe: Link $l = \{(c_1, c_2)\}$

```
1: function C1C2INCOMINGLINKSFEATURE(L)
2:     occ = 0
3:     for all a_i ∈ c_2 do
4:         if c_i hat ein Wikilink zu einem Artikel a_j ∈ c_1 then
5:             occ = occ + 1
6:     return occ
```

Ausgabe: Eine natürliche Zahl

Die Features `c1distanceCommonAncestorFeature` und `c2distanceCommonAncestorFeature` berechnen den Abstand zwischen den gegebenen Kategorien c_1 und c_2 zum ersten gemeinsamen Vorfahren c_A beider Kategorien. Die Entfernungen werden einzeln berechnet, d.h. `c1distanceCommonAncestorFeature` berechnet den Abstand von c_1 zu c_A und `c2distanceCommonAncestorFeature` den Abstand c_2 zu c_A. Ein Beispiel wird in Abbildung 40 gezeigt. Die Berechnung der Distanz findet durch einen adaptierten Breitensuche-Algorithmus [32] statt. Es werden zwei Breitensuchen im Kategoriengraph ausgehend von c_1 und c_2 gestartet und besuchte Knoten gespeichert. Die Suche erfolgt solange, bis sich die Wege an einem einzigen Knoten kreuzen. Der Algorithmus wird in 5.2.9 angegeben.

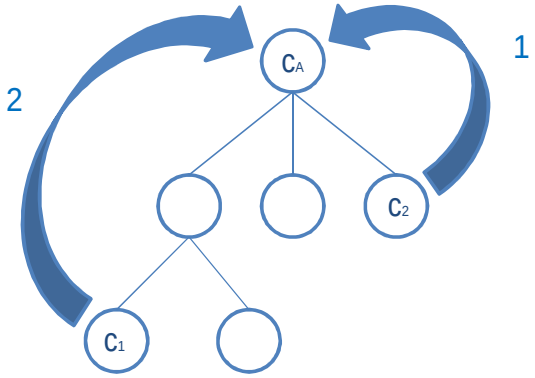

Abbildung 40: Entfernung von zwei Knoten zum ersten gemeinsamen Vorfahren

Die Features `c1NumberOfSubcategories`, `c1NumberOfSupercategories`, `c2NumberOfSubcategories` und `c2NumberOfSupercategories` zählen die Anzahl der Unter- und Oberkategorien von c_1 und c_2. Kategorien mit einer Vielzahl von Unterkategorien repräsentieren in der Regel abstraktere Konzepte, die von vielen anderen Artikeln referenziert werden, wie z.B. „Wissenschaft“. Schließlich zählt das Feature `CommonWikilinksFeature` die Anzahl der gemeinsamen Wikilinks zwischen c_1 und c_2.

Algorithmus 5.2.9 Berechnung des letzten gemeinsamen Vorfahren

Eingabe: Link $l = \{(c_1, c_2)\}$

Prozedur DISTANCETOCOMMONANCESTOR(L)
▷ Initialisierung
$paths_1 = \{\}$
$paths_2 = \{\}$
$visitednodes_1 = \{\}$
$visitednodes_2 = \{\}$
▷ Aktuelle Kategorie hinzufügen
$path_1 = \text{ADD\_CATEGORY\_TO\_PATH}(c_1, paths_1)$
$path_2 = \text{ADD\_CATEGORY\_TO\_PATH}(c_2, paths_2)$
$visitednodes_1 = visitednodes_1 \cup c_1$
$visitednodes_2 = visitednodes_2 \cup c_2$
repeat
$tmp_1 = \text{GET\_PARENTS}(c_1, paths_1)$
$tmp_2 = \text{GET\_PARENTS}(c_2, paths_2)$
for all $c_{1,i} \in tmp_1$ **do**
$path_1 = \text{ADD\_CATEGORY\_TO\_PATH}(c_{1,i}, paths_1)$
$visitednodes_1 = visitednodes_1 \cup c_{1,i}$
for all $c_{2,i} \in tmp_2$ **do**
$path_2 = \text{ADD\_CATEGORY\_TO\_PATH}(c_{2,i}, paths_2)$
$visitednodes_2 = visitednodes_2 \cup c_{2,i}$
until $\exists c_A \in visitednodes_1 \cap visitednodes_2$
$distance_1 = \text{CALCULATE\_DISTANCE\_TO\_NODE}(c_A, paths_1)$
$distance_2 = \text{CALCULATE\_DISTANCE\_TO\_NODE}(c_A, paths_2)$

Ausgabe: Zwei Werte $distance_1$ und $distance_2$, die den Werten der Features entsprechen

5.2.1.4 *Artikelbasierte Features*

Diese Menge von Features bestimmt sich aus dem Inhalt der Wikipedia-Artikel. Das erste Feature macht sich die Tatsache wieder zunutze, dass der erste Satz eines Artikels eine Definition des Konzeptes enthält. `definitionSentenceFeature` gibt an, ob im ersten Satz des Artikels a_1 aus Kategorie c_1 der lexikalische Kopf von Kategorie c_2 vorkommt. Beispielsweise liefert dieses Feature für den Link $L = (Mäuseartige, Nagetiere)$ den Wert 0 zurück, weil „Nagetiere" im ersten Satz des Artikels zu „Mäuseartige" vorkommt. Auch hier gilt, dass keine sprachabhängige Suche von Mustern benötigt wird, und das Feature somit in verschiedenen Sprachen angewendet werden kann.

Das Feature *c2Inc1Feature* zählt die Anzahl von Vorkommen der lexikalischen Köpfe von c_2 im Rest des Artikels von c_1. Weiterhin geben `c1ArticleFeature` und `c2ArticleFeature` an, ob es überhaupt einen Artikel mit den Namen von c_1 bzw. c_2 gibt. Dass ein Artikel zu einer Kategorie gibt, ist ein Hinweis darauf, dass es sich bei dieser Kategorie um ein Konzept handelt. Sie liefern den Wert 1 zurück, falls ein Artikel zu einer Kategorie existiert, ansonsten den Wert 0. Zum Schluss gibt `RedirectFeature` an, ob der Artikel a_1, der die Kategorie c_1 beschreibt, zum entsprechenden Artikel a_2 weiterleitet, der die Kategorie c_2 beschreibt. Dies kann ein Hinweis darauf sein, dass eine Synonymie-Relation zwischen c_1 und c_2 besteht.

5.2.2 *Sprachunabhängigkeit des Verfahrens*

Der hier vorgestellte Ansatz (TaxWikiML.KOM) ist genauso wie TaxWikiHeur.KOM sprachunabhängig, da der Ansatz auf verschiedene Sprachen ohne Modifizierung der Features angewendet werden kann. Der Ansatz kann benutzt werden, um eine große Menge an taxonomischen Relationen aus der Menge der Links im Wikipedia-Kategoriengraph zu extrahieren. Dazu werden nur folgende wenige Informationen benötigt:

1. Eine Liste von Präfixen, die Wikipedia benutzt, um Metakategorien zu erkennen, wie z.B. „Wikipedia" oder „Benutzer"(siehe Abschnitt 5.2.1.1).
2. Die Präposition, die in der Sprache S benutzt wird, um Verfeinerungslinks zu definieren. Beispielsweise „by" für Englisch oder „nach" für Deutsch (siehe Abschnitt 5.2.1.1).
3. Die Liste aller Präpositionen in Sprache S, um lexikalische Köpfe in komplexen Kategorien heuristisch zu bestimmen (siehe Abschnitt 5.2.1.2).

Im Vergleich zum ersten Ansatz braucht TaxWikiML.KOM keine Angabe darüber, wo der lexikalische Kopf in der jeweiligen Sprache zu finden ist. Mit Hilfe von diesen Informationen ist es möglich, Hyponymie-Beziehungen zu erkennen. Die Güte des Verfahrens in verschiedenen Sprachen wird im nächsten Abschnitt dargestellt.

5.2.3 *Evaluation des Verfahrens*

TaxWikiML.KOM wurde in fünf verschiedenen Sprachen evaluiert: drei europäische Sprachen (Englisch, Deutsch und Spanisch) und zwei Sprachen mit nicht-lateinischen Schriften (Arabisch, Russisch). Die Evaluation bestand aus drei Schritten:

1. Erstellung von manuell gelabelten Korpora für die fünf verschiedenen Sprachen (Abschnitt 5.2.3.1).

2. Überprüfung der Klassifikationsergebnissen mit Hilfe der manuell gelabelten Korpora (Abschnitt 5.2.3.2)

3. Vergleich zu externen Wissensbasen (WordNet, WikiNet) bei der Klassifikation aller Links des Wikipedia-Kategoriengraphs (vgl. Abschnitt 5.2.3.3)

Im Grunde handelt es sich um dieselbe Evaluationsmethodologie, mit der TaxWikiHeur.KOM im Abschnitt 5.1.4 evaluiert wurde. Der wichtigste Unterschied ist, dass beim Vergleich mit anderen Wissensbasen nicht nur manuell erstellte Wissensbasen betrachtet wurden, sondern auch der WikiNet-Ansatz, der auf Interwikilinks setzt [106]. Darüber hinaus wird jetzt TaxWikiML auch für die russische Sprache evaluiert.

Bei der Evaluation von TaxWikiML.KOM hat sich die Frage gestellt, ob die manuell erstellten Korpora aus Abschnitt 5.1.4.1 wieder benutzt werden sollten. Diese Frage wurde verneint, weil der vorherige Korpus den Kategoriengraphen nicht gut abgebildet hat: Eine der Forschungsziele von TaxWikiHeur.KOM war es, zu beweisen, dass eine multilinguale automatische Klassifikation von Links möglich ist. Bei der Evaluation ging es darum, eine möglichst heterogene Menge von Links zu bekommen, um die Nutzbarkeit des Verfahrens nachzuweisen. Dieses Ziel wurde erreicht, aber mit dem Nachteil, dass die Links mehrheitlich aus dem unteren Bereich des Kategoriengraphs stammten. Ponzetto et al. haben in [117] festgestellt, dass sich die Struktur des Kategoriengraphen zwischen den oberen und den unteren Bereichen unterscheidet. Dies bedeutet, dass die vorher erstellten Korpora den Kategoriengraphen nicht in seiner Gesamtheit repräsentiert. Da aber beim maschinellen Lernen der Trainingskorpus eines Klassifikators besonders (vgl. Abschnitt 2.2.2) wichtig ist, wurde ein neuer Korpus nötig.

5.2.3.1 *Erstellung des Evaluationskorpus*

Zur Bestimmung des Korpus wurden 1000 zufällig gewählte Wikipedia-Kategorien verwendet. Die Kategorien wurden mit Hilfe der „zufälliger Artikel"-Funktion[7] der Wikipedia nach Algorithmus 5.2.10 extrahiert:

7 `http://de.wikipedia.org/wiki/Special:Random` - Zugriff am 14.11.2012

Algorithmus 5.2.10 Extraktionsalgorithmus für die Korpora

Eingabe: Wikipedia-Artikel $A = a_1,...,a_n$, Wikipedia-Kategorien $C = c_1,...,c_m$

Prozedur CORPUS-EXTRACTOR(A,C)
  $O = \{\}$ ▷ Initialisierung Ausgabemenge
  a_i = GET_RANDOM_ARTICLE(A)
  $O = O \cup$ GET_ALL_CATEGORIES(a_i)
  c_i = GET_RANDOM_CATEGORY(a_i)
  $O = O \cup$ GET_CATEGORIES(c_i) ▷ Wähle nur Kategorien, die weniger als 100 Oberkategorien haben, um eine Überrepräsentation einer bestimmten Domain zu vermeiden
  repeat
    c_j = GET_RANDOM_CATEGORY(c_i)
    $O = O \cup$ GET_CATEGORIES(c_j)
    $c_i = c_j$
  until $|O| < 1000$

Ausgabe: Menge von zufällig ausgewählten Kategorien O

Ausgehend von einem zufällig gewählten Artikel wird ein zufällig gewählter Pfad bis zur Wurzel verfolgt. Die in diesem zufälligen Pfad vorkommenden Kategorien werden gespeichert und später heruntergeladen. Anschließend wird ein neuer zufälliger Artikel gewählt und der Prozess nochmal durchgeführt. Dies wird so lange wiederholt, bis sich 1000 Kategorien im Korpus befinden.

Die Korpusextraktion geht wie in 5.1.4.1 weiter: Die gespeicherten Kategorien und ihre Oberkategorien werden als Links im CSV-Format gespeichert. Abschließend werden alle gefundenen Links manuell mit ihrer Klasse (*is-a* oder *not-is-a*) gespeichert. Ein Überblick der extrahierten Korpora wird in Tabelle 9 dargestellt.

Tabelle 9: Überblick über die extrahierten Korpora

Sprache	Englisch	Spanisch	Deutsch	Arabisch	Russisch
Anzahl *is-a* links	1293	786	808	1135	2545
Anzahl *not-is-a* links	3043	1388	1597	1604	3572
Gesamtanzahl Links	4336	2174	2405	2739	6297

In den extrahierten Korpora ist insgesamt ein höherer Anteil von *not-is-a*-Links als *is-a*-Links festzustellen, allerdings war das Verhältnis in den verschiedenen Sprachen sehr unterschiedlich. Die Werte gehen von 56 % *not-is-a*-Links für Russisch bis zu 56 % *not-is-a*-Links bei der englischen Sprache (siehe Tabelle 9). Das könnte z.B. darauf hinweisen, dass die englische Wikipedia zu einer „Überkategorisierung" neigt, während in anderen Sprachen eine konsequentere Kategorisierung realisiert wird [52].

5.2.3.2 *Überprüfung der Klassifikationsergebnisse mit Hilfe der manuell gelabelten Korpora*

In diesem Abschnitt werden die Ergebnisse der Klassifikation der zum Korpus gehörenden Links präsentiert. Für die Evaluation wurden das Weka Machine Learning Toolkit [51] und J48-Entscheidungsbäume (Weka-Umsetzung von C4.5-Bäumen [120])

als Klassifikator ausgewählt. Die zuvor vorgestellten Features wurden für alle im Korpus vorhandenen Links berechnet und verwendet, um den Klassifikator zu trainieren. Entscheidungsbäume sind ein häufig verwendeter Klassifikator, da sie sich sowohl schnell trainieren lassen als auch schnell beim Klassifizieren von Instanzen sind. Darüber hinaus sind die Entscheidungsregeln für Menschen verständlich und sie lassen sich mit anderen Entscheidungstechniken kombinieren, um die Ergebnisse zu verbessern (wie z.B. mit einer Kostenmatrix wie im Abschnitt 5.2.3.3 geschehen).

Alle Klassifikationsergebnisse wurden einer zehnfachen stratifizierte Kreuzvalidierung unterzogen (siehe Abschnitt 2.2.4). Tabelle 10 gibt einen Überblick über die Ergebnisse. Sie zeigt richtig und falsch klassifizierte Instanzen. Im Durchschnitt über alle Sprachen werden 83,1 % der Links korrekt und 16,9 % falsch markiert.

Tabelle 10: Zusammenfassung der Ergebnisse nach Sprachen

Sprache	Englisch	Spanisch	Deutsch	Arabisch	Russisch
Korrekt klass. Inst.	3590	1838	1963	2283	5067
Inkorrekt klass. Inst.	746	336	442	456	1230
Gesamtanzahl der Links	4336	2174	2405	2739	6297

Tabelle 11 fasst die Ergebnisse der Metriken zur Bewertung von Klassifikationsalgorithmen zusammen: Precision, Recall und F_1-Maß. Für Precision lagen die Ergebnisse über alle Sprachen durchschnittlich bei 74,5%, während der Recall 77,2 % für *is-a*-Relationen betrug. Für *not-is-a*-Relationen betrugen Precision 87% und Recall 86,3%.

Tabelle 11: Precision, Recall und F_1-Maß für jede Klasse und Sprache

	Precision	Recall	F_1-Maß	Klasse
Englisch	72,5 %	68,1 %	70,3 %	is-a
	86,8 %	89,0 %	87,9 %	not-is-a
Spanisch	76,3 %	83,1 %	79,5 %	is-a
	89,9 %	85,4 %	87,6 %	not-is-a
Deutsch	71,9 %	74,4 %	73,1 %	is-a
	86,8 %	85,3 %	86,0 %	not-is-a
Arabisch	79,7 %	80,4 %	80,0 %	is-a
	86,0 %	85,5 %	85,7 %	not-is-a
Russisch	73,2 %	81,5 %	77,1 %	is-a
	86,4 %	79,8 %	83,0 %	not-is-a

Die Ergebnisse der Evaluation wurden mit Hilfe der Konfusionsmatrix analysiert. Für Englisch wird die Konfusionsmatrix beispielhaft in Tabelle 12 dargestellt. Diese Matrix zeigt die größte Fehlerquelle besonders gut. Es handelt sich um falsch klassifizierte *is-a*-Links. Der Grund dafür ist, dass eine hohe Anzahl von *is-a*-Links nicht durch einzelne Features erkannt werden können, sondern durch Kombination von mehreren Features. Diese Kombinationen können mehr Instanzen als einfache Features erkennen, bringen aber eine weitere Ungenauigkeit mich sich.

Tabelle 12: Konfusionsmatrix für die englische Sprache

a	b	← klassifiziert als
944	349	a = is-a
404	2639	b = not-is-a

Zusätzlich wurden die Beiträge der einzelnen Featuretypen auf die Performanz (Vertrauenswahrscheinlichkeit) evaluiert. In Tabelle 13 werden die Ergebnisse pro Klasse gezeigt. Man stellt fest, dass die einzelnen Feature-Typen in den meisten Fällen nicht mehr als 75% der Links korrekt klassifizieren.

Tabelle 13: Performanz der Features pro Sprache

	Vorverarbeitung	Syntaktisch	Strukturell	Artikelbasiert
Englisch	70,2 %	69,8 %	75,8 %	71,4 %
Spanisch	63,8 %	69,9 %	73,6 %	70,5 %
Deutsch	66,4 %	68,1 %	74,1 %	67,2 %
Arabisch	64,2 %	59,8 %	77,8 %	73,5 %
Russisch	66,2 %	64,6 %	69,1 %	66,8 %

Schließlich wurde der Informationsgehalt der einzelnen Features untersucht. Dies wurde mit Hilfe des Maßes *Informationsgewinn* (engl. Information Gain) [103] gemessen. Der Informationsgewinn gibt an, wie gut ein Feature die Instanzen entsprechend ihrer Klasse trennt und somit die Entropie [103] reduziert. Ein Überblick der Ergebnisse wird in Tabelle 14 gezeigt. Die Werte in der Tabelle entsprechen den in Tabelle 8 vergebenen IDs der Features. Features, die benutzt werden, um *not-is-a*-Links zu erkennen, werden höher gerankt als Features, die benutzt werden, um *is-a*-Relationen zu erkennen, da mehr *not-is-a*-Links als *is-a*-Links existieren. Dadurch sind sie für den Informationsgewinn aussagekräftiger. Syntaktische und strukturelle Features schnitten am besten ab. Außerdem lässt sich sagen, dass strukturelle Features am besten zur Erkennung von *not-is-a*-Links und syntaktische Features von *is-a*-Links benutzt werden können.

Darüber hinaus ließ sich beobachten, dass die Features `distanceC1ToCommonAncestor` und `distanceC2ToCommonAncestor` keine große Unterscheidungskraft hatten. Dies kann durch die Methode zur Korpus-Extraktion (Abschnitt 5.2.3.1) erklärt werden. Im Korpus wurden nur direkte Links zwischen Kategorien verwendet. Allerdings könnten diese Features in anderen Szenarien hilfreich sein, z.B. um einen Klassifikator zu trainieren, der nicht nur direkte, sondern auch indirekte (transitive) Links korrekt klassifizieren kann. Ein solcher Klassifikator könnte benutzt werden, um *is-a*-Relationen unabhängig vom Einsatzszenario dieser Arbeit zu erkennen. Dies kann Teil zukünftiger Forschung sein.

Tabelle 14: Ranking der benutzten Features für jede Sprache (IDs der Features stehen in Tabelle 8)

	Englisch	Spanisch	Deutsch	Arabisch	Russisch
1.	18	7	7	18	3
2.	2	18	2	19	7
3.	20	3	19	6	18
4.	7	20	20	16	2
5.	8	2	3	2	16

5.2.3.3 *Vergleich zu externen Wissensbasen*

TaxWikiML.KOM wurde zusätzlich auf den gesamten Kategoriengraph der Wikipedia angewendet. Die Ergebnisse wiesen einen hohen Anteil von falsch-positiv klassifizierten Instanzen auf. Das bedeutet, dass zu viele Links als *is-a* gekennzeichnet wurden, wodurch viele *not-is-a*-Links unnötig als *is-a* markiert wurden. Das führte zu einem Recall (93%) bei *is-a*-Instanzen, die Precision (60.5%) war aber nicht so hoch. Insgesamt war die Performanz des Verfahrens nicht höher als 70% (siehe Tabelle 15). Die detaillierten Ergebnisse (u.a. Precision und Recall) werden im Anhang A.2 dargestellt.

Tabelle 15: Zusammenfassung der Ergebnisse ohne Optimierung

Anzahl der korrekt klassifizierten Instanzen	28309 (68,75%)
Anzahl der inkorrekt klassifizierten Instanzen	12868 (21,25%)
Gesamtanzahl der Instanzen	41177

Im maschinellen Lernen werden Klassifikationsergebnisse vom Trainingkorpus und vom Klassifikator selbst beeinflusst [167]. Ein Klassifikator könnte, z.B. durch einen im Vergleich zu realen Daten nicht repräsentativen Trainingskorpus, so gewichtet werden, dass eine Klassifikation von realen Daten nicht so gut wie die Klassifikation mit Trainingsdaten ist. Um diesem Umstand entgegenzuwirken, wurde ergänzend eine Kostenmatrix definiert und benutzt, um den Klassifikator anzupassen. Mit Hilfe einer Kostenmatrix können Kosten für bestimmte Fehler erhöht bzw. reduziert werden. Im Fall von TaxWikiML.KOM können entweder falsch-positive Fehler (*not-is-a* als *is-a* markiert) oder falsch-negative Fehler (*is-a* als *not-is-a* markiert) höher bestraft werden. Der Klassifikator wird dann neu erstellt, so dass neben der zuverlässigen qualitativen Klassifikation auch die angefallenen Fehlerkosten berücksichtigt werden. Dadurch können unerwünschten Fehler reduziert werden. Es handelt sich um ein „trade-off“, denn bei einem zu hohen Kostenfaktor verschlechtern sich andere Fehlertypen. Aus diesem Grund wurde der optimale Kostenfaktor auf empirischem Weg ermittelt.

In diesem Abschnitt werden die Ergebnisse dieser „optimierten Version“ von TaxWikiML.KOM vorgestellt. Folgende Kostenmatrix wurde bspw. für Englisch gewählt:

$$\begin{pmatrix} 0 & 18 \\ 1 & 0 \end{pmatrix}$$

Die Kostenmatrix gibt an, dass falsch-positive Fehler viel stärker gewichtet werden als andere Fehler (Faktor 18).

TaxWikiML.KOM wird für die englische Sprache auf den gesamten Kategoriengraph der Wikipedia (mit der oben genannten Kostenmatrix) angewendet und die Genauigkeit wird mit Hilfe von WordNet und WikiNet verglichen. WikiNet wurde benutzt, weil es sich um einen vergleichbaren Ansatz handelt, der auf Interwikilinks basiert (also dem von Wikipedia zur Verfügung gestellten Konzept für solche Aufgaben). Außerdem greift WikiNet auf keine weiteren externen Quellen zu. Die englische Sprache wurde gewählt, weil Englisch die Basis-Sprache für WikiNet ist und alle anderen Relationen werden über die Interwikilinks der Wikipedia inferiert. Darüber hinaus gilt die englische Wikipedia heutzutage als größte und wichtigste Wikipedia-Version. Die Ergebnisse von WikiTaxML.KOM für Deutsch im Vergleich zu WikiNet waren sehr ähnlich und ließen keine eindeutige Aussage zu, welches Verfahren besser abschneidet. Die Ergebnisse stehen der Vollständigkeit halber im Anhang A.2.2.

Der Vergleich für Englisch findet genau wie bei der Evaluation von TaxWikiHeur.KOM statt: Es wurden Links ausgewählt, die sowohl in WordNet und WikiNet vorkommen (siehe Abbildung 41).

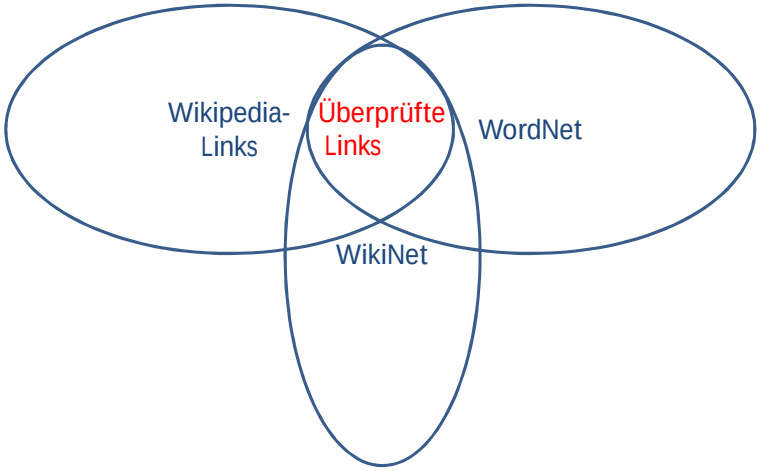

Abbildung 41: Links, die sowohl im Kategorien-Graph als auch in WordNet und WikiNet vorkommen

Insgesamt wurden für die englische Sprache 15.483 Links überprüft, die sowohl im Kategoriengraph als auch bei WordNet und WikiNet vorkommen. Durch die Anzahl der korrekt und falsch klassifizierten Links lassen sich Evaluationsmaße wie Precision, Recall und F_1-Maß für TaxWikiML.KOM und WikiNet berechnen und vergleichen.

Tabelle 16 gibt einen Überblick über die Ergebnisse. Sie zeigt korrekt klassifizierte Links von TaxWikiML.KOM und WikiNet. 85,95% der Links wurden von TaxWikiML.KOM korrekt klassifiziert. Für WikiNet waren es 78,23% der Links.

Tabelle 16: Vergleich der Ergebnisse zwischen TaxWikiML.KOM und WikiNet

	TaxWikiML.KOM	WikiNet
Korrekt klass. Links	13307 (85,95%)	12113 (78,23%)
Inkorrekt klass. Links	2176 (14,05%)	3370 (21,77%)
Gesamtanzahl der Links	15483	15483

Tabelle 17 zeigt detaillierte Ergebnisse beider Ansätze für den Evaluationskorpus. Für F_1-Maß lagen die Werte für *is-a*-Beziehungen bei 80,48% und bei 89,02% für not-is-a Beziehungen. Für die englische Sprache lässt sich also sagen, dass TaxWikiML.KOM auf dem Evaluationskorpus besser als WikiNet arbeitete. Leider konnten nur 15.483 von 85.938 Links evaluiert werden. Dafür gibt es hauptsächlich zwei Gründe: Zum einen hat Wikipedia eine viel größere Konzept-Abdeckung als WordNet und zum anderen können viele Wikipedia-Kategorien (z.B. „Das große Buch der deutschen Fußballvereine“) nicht einem WordNet-Synset zugeordnet werden.

Tabelle 17: Detaillierte Ergebnisse von TaxWikiML und WikiNet im Vergleich zu WordNet

	Precision	Recall	F_1-Maß	Klasse
TaxWikiML.KOM	86,12 %	73,54 %	80,48 %	is-a
	85,86 %	92,42 %	89,02 %	not-is-a
WikiNet	69,14 %	78,29 %	73,37 %	is-a
	85,20 %	78,29 %	81,60 %	not-is-a

5.3 ZUSAMMENFASSUNG

In diesem Kapitel wurden zwei Verfahren vorgestellt, mit deren Hilfe Hyponymierelationen zwischen Kategorien in Wikipedia erkannt werden können. Im Gegensatz zu anderen Verfahren beruht die Erstellung der Taxonomie weder auf sprachabhängigen Methoden noch auf bereits existierenden (manuell erstellten) Wissensbasen. Beim ersten Verfahren handelt es sich um ein regelbasiertes Verfahren: TaxWikiHeur.KOM, das in der Lage ist, zwischen *is-a*- und *not-is-a*-Relationen zu unterscheiden. Mit diesem Verfahren wurde gezeigt, dass die multilinguale automatische Erkennung von *is-a*-Relationen möglich ist. TaxWikiHeur.KOM wurde evaluiert und die Ergebnisse wurden mit einem manuell erstellten Korpus sowie mit zwei von Experten erstellten Wissensbasen verglichen. Darüber hinaus wurden die Grenzen des Ansatzes gezeigt und analysiert.

Auf der Basis der bei TaxWikiHeur.KOM gewonnenen Erkenntnisse wurde ein weiterer Ansatz, TaxWikiML.KOM, entwickelt. TaxWikiML.KOM besteht aus einem trainierten binären Klassifikator, der mit Hilfe einer Reihe von Features automatisch taxonomische Beziehungen zwischen Paaren aus dem Wikipedia-Kategoriengraph erkennen kann. Das Verfahren zeichnet sich dadurch aus, dass man mit wenigen Informationen über eine Zielsprache und ohne externe Quellen eine Klassifikation

von Links vornehmen kann. Wie im ersten Fall wurde die Evaluation mit Hilfe von manuell und automatisch erstellten Korpora durchgeführt.

Beide Verfahren ermöglichen die Ableitung von Taxonomien aus Wikipedia für verschiedene Sprachen durch Benutzung von syntaktischen und strukturellen Regeln. Im Rahmen dieser Arbeit wurden diese Verfahren für die Erstellung von Taxonomien zum Einsatz im Ressourcen-basierten Lernen in Online Communities benutzt. Allerdings sind auch andere Szenarien denkbar: Sie kann z.B. benutzt werden, um weitere automatisch erstellte Wissensbasen zu evaluieren. Insbesondere für Sprachen, wo keine manuell erstellten Wissensbasen zum Vergleich existieren oder nicht verfügbar sind.

Im nächsten Kapitel wird detailliert erklärt, wie die erstellte Taxonomie in der CROKODIL-Plattform eingesetzt wird.

6

IMPLEMENTIERUNG UND PROOF-OF-CONCEPT

»The classification of facts and the formation of absolute judgments upon the basis of this classification [...] essentially sum up the aim and method of modern science.«

— Karl Pearson

In Kapitel 4 wurden bereits die Funktionen der CROKODIL- Plattform und das im Rahmen dieser Arbeit entwickelte Konzept zur Steigerung der Zugreifbarkeit auf Ressourcen im Ressourcen-basierten Lernen in online Communities durch die Verwendung von Taxonomien beschreiben. In diesem Kapitel wird zunächst in Abschnitt 6.1 die bestehende Architektur der CROKODIL-Plattform sowie die in dieser Arbeit vorgenommene Erweiterung der Architektur vorgestellt. Abschnitt 6.2 beschreibt die im Rahmen dieser Arbeit implementierte Erweiterung des CROKODIL-Datenmodells sowie die zusätzliche Realisierung von Ressourcenempfehlungen. Die verschiedenen implementierten Erweiterungen zum Export der von den Benutzern in CROKODIL verwendeten Tags, die Prüfung auf Hyponomiebeziehungen zwischen diesen, die mittels einer mit Hilfe des im vorhergehenden Kapitel beschriebenen Verfahrens bestimmten Taxonomie-Datenbank erfolgt, sowie die Integration der gefundenen Relationen in das CROKODIL-Modell erläutert Abschnitt 6.3.

6.1 CROKODIL-KOMPONENTEN UND ERWEITERTE ARCHITEKTUR

Die Architektur der CROKODIL-Plattform, inkl. der Empfehlungssysteme, besteht aus verschiedenen Komponenten, die in Abbildung 42 dargestellt sind. In einer semantischen Datenbank werden die der CROKODIL-Plattform zu Grunde liegenden semantischen Informationen gespeichert. Das Web-Portal und der Net-Navigator erlauben dem Benutzer die Manipulation der Plattformindividuen und die Navigation im semantischen Netz innerhalb seines Browsers. Web-Services dienen zur Kommunikation mit dem Backend und damit zur Anbindung weiterer Dienste.

Im Rahmen dieser Arbeit wurden ergänzend eine Taxonomiedatenbank realisiert, welche die mit Hilfe des in Kapitel 5 vorgestellten Verfahrens TaxWikiML.KOM klassifizierten Beziehungen zwischen den Wikipedia-Kategorien enthält, und verschiedene Tools, CrokoTaxTools genannt, die zur Manipulation der Datenbank und Anbindung an CROKODIL dienen. Sie werden detailliert in Abschnitt 6.3 dargestellt.

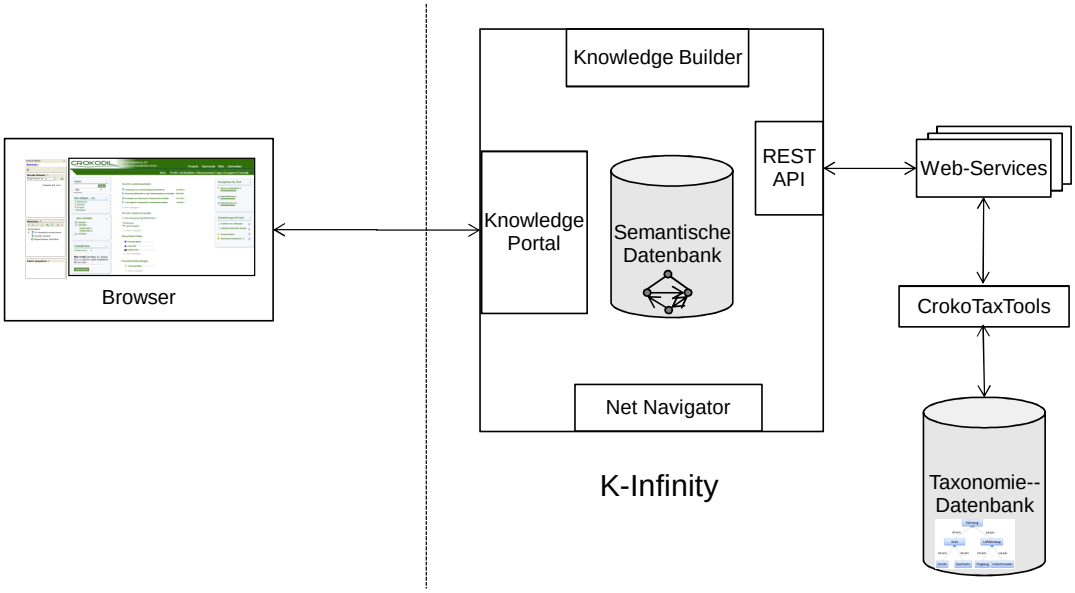

Abbildung 42: Gesamtarchitektur der CROKODIL-Plattform

6.1.1 *CROKODIL-Komponenten*

CROKODIL basiert auf K-Infinity, was eine Technologieplattform der Firma intelligent views[1] zur Vernetzung und Verwaltung von Wissen in Wissensnetzen ist. K-Infinity und damit auch die CROKODIL-Plattform setzt sich aus den folgenden Komponenten zusammen:

- einer semantischen Datenbank, die das Wissensnetz speichert,
- Werkzeugen zur Modellierung des Wissensnetzes und zur Definition von sogenannten Expertensuchen
- Schnittstellen für den Zugriff auf das Wissensnetz durch externe Anwendungen,
- einem Web-Portal als Frontend für die Visualisierung und die Bearbeitung von Entitäten des Wissensnetzes durch die Nutzer,
- einer graphischen Visualisierung und Navigationsschnittstelle für das semantische Netz (Net-Navigator).

K-Infinity ist plattformunabhängig, da die Komponenten in Java und Smalltalk implementiert wurden. Während das Web-Portal das Web-Interface der CROKODIL-Plattform ist, wird eine *Knowledge Builder* genannte Applikation zur Modellierung von Wissensnetzen genutzt. Mit dem Knowledge Builder können Wissensnetzknoten (im Fall von CROKODIL sind das Ressourcen, Tags, Aktivitäten usw.) und Relationen erzeugt, bearbeitet und gelöscht werden. Wissensnetzknoten werden in K-Infinity u.a. in Begriffe und Individuen unterteilt. Für jeden Wissensnetzknoten (Begriffe und Individuen) können Attribute und Relationen definiert werden. Während die Begriffe ähnlich wie Klassen gemeinsame Attribute von Individuen zusammenfassen, bilden die Individuen den tatsächlichen Inhalt des Wissensnetzes.

1 `http://www.intelligent-views.com` - Zugriff am 14.11.2012

Mittels des Knowledge Builders lassen sich außerdem die sogenannten *Expertensuchen* definieren. Eine Expertensuche erlaubt die Zusammenstellung von Begriffen, Individuen, Relationen und Attributen zu geschachtelten, feststehenden Suchanfragen, deren Ergebnisse sich dynamisch ändern, wenn neue Individuen oder Relationen angelegt oder gelöscht werden. Die Ergebnisse der Expertensuchen können im Portal in eigenen Bereichen angezeigt werden. Abbildung 43 zeigt die Definition einer Expertensuche in der CROKODIL-Plattform. Die dargstellte Expertensuche in Abb. 43 liefert alle Themen von Ressourcen, die ein gemeinsames Thema mit einer gegebenen Ressource teilen.

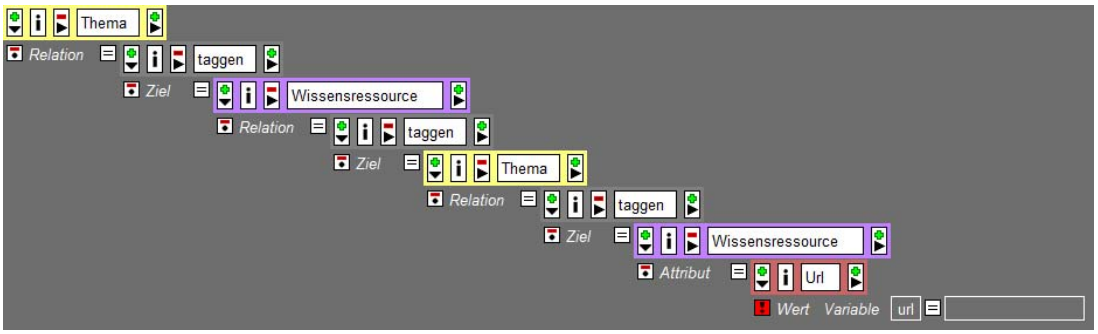

Abbildung 43: Eine einfache Expertensuche im Knowledge Builder

Ein wichtiger Vorteil von Expertensuchen ist die Tatsache, dass die Ergebnisse Mengen sind und K-Infinity Mengenoperatoren anbietet, mit denen sich Expertensuchen beliebig kombinieren lassen. Beispielsweise lassen sich die Vereinigung und die Schnittmenge von Expertensuchen bilden.

Das graphische Anzeigen der Wissensnetze übernimmt der *Net-Navigator*. Wie der Name schon andeutet, kann der Net-Navigator die Wissensnetze nicht nur anzeigen, sondern erlaubt das Navigieren durch ein Wissensnetz. Die Objekte eines Wissensnetzes und ihre Verbindungen miteinander werden als Graph dargestellt. Abbildung 44 zeigt einen Ausschnitt aus einem Wissensnetz im Net-Navigator. Ein Benutzer „renato“ besitzt eine Ressource über Web-Didaktik. Darüber hinaus teilen sich „renato“ und „Christoph R.“ einen Tag „Ressourcen-basiertes Lernen“ vom Typ Thema. Benutzer „Christoph R.“ hat eine Ressource „PERKAM“ mit dem Thema „context-based learning“ getaggt.

K-Infinity bietet die Möglichkeit, Web-Services zu definieren, die mittels einer REST-API[2] den Zugriff auf die Backend-Datenbank erlauben. REST-Services können über URLs[3] adressiert angesprochen werden. Eine direkte Manipulation von Objekten ist nicht vorgesehen. Jeder Zugriff muss indirekt über die dem Objekt zugeordnete URL erfolgen. Eine zentrale Bedeutung bei REST haben die HTTP-Methoden GET, POST und DELETE. Sie stellen die Funktionen dar, die auf die Objekte angewendet werden können. GET steht für den Aufruf von Informationen, POST für das Einfügen von Informationen und DELETE für das Löschen von Informationen. Für eine komplette Beschreibung von REST wird an dieser Stelle auf [132] verwiesen.

Mit Hilfe der REST-API lassen sich also die Informationen, d.h. Objekte und Relationen, abrufen. In Abschnitt 6.3 wird die Nutzung der Web-Services zur Realisierung der auf Taxonomie-basierenden Empfehlungen erläutert.

2 REpresentational State Transfer Architektur (REST)
3 Uniform Resource Locator (URL)

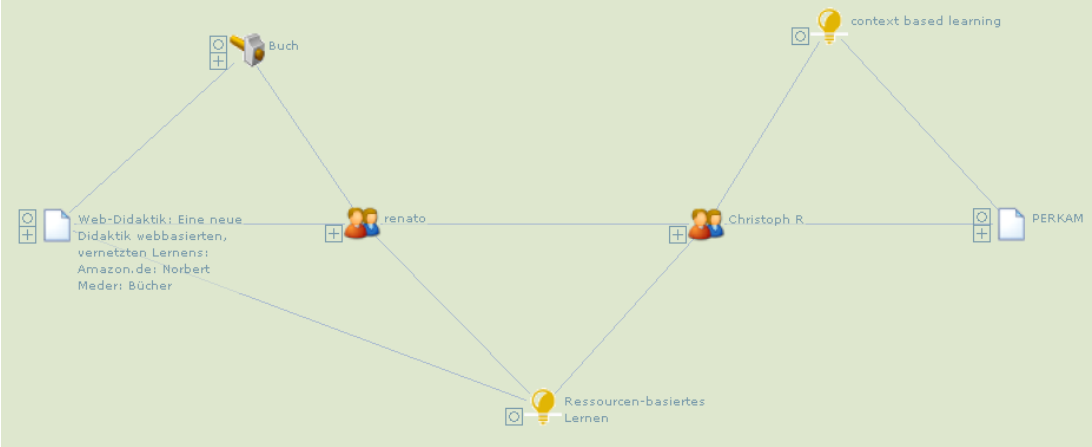

Abbildung 44: Ausschnitt eines Wissensnetzes im Net-Navigator

6.1.2 *Die Taxonomiedatenbank*

Die Taxonomiedatenbank ist eine SQL-Datenbank, die alle aus der Wikipedia relevanten Informatioonen enthält, die für TaxWikiML.KOM notwendig sind sowie die von TaxWikiML.KOM klassifizierten Links zwischen den Wikipedia Kategorien enthält. Der Wikipedia-Korpus wurde als XML-Dump von der Wikipedia-Dump-Seite[4] heruntergeladen. Die notwendigen Informationen wurden in einem Vorbereitungsschritt in eine MySQL-Datenbank überführt. Die relevanten Tabellen für die deutsche Sprache werden in Abbildung 45 dargestellt. PS steht dabei für Primärschlüssel und FS für Fremdschlüssel. Jede der Tabellen enthält spezifische Informationen: `de_lemma` enthält alle Kategorien und Artikel mit ihren jeweiligen IDs, `de_article` enthält zusätzlich den Inhalt der Artikel, `de_wikilink` stellt die Wikilinks dar, also Verweise von einem Artikel zu einem anderen, `de_redirect` enthält die Weiterleitungen von einem Artikel zu einem anderen und schließlich `de_category`, die den Kategoriengraph darstellt. Die Tabelle `de_category` enthält eine zusätzlichen Spalte, die mittels WikiTaxML.KOM berechnet wurde und die angibt, ob es zwischen zwei Kategorien eine Hyponymie gibt oder nicht.

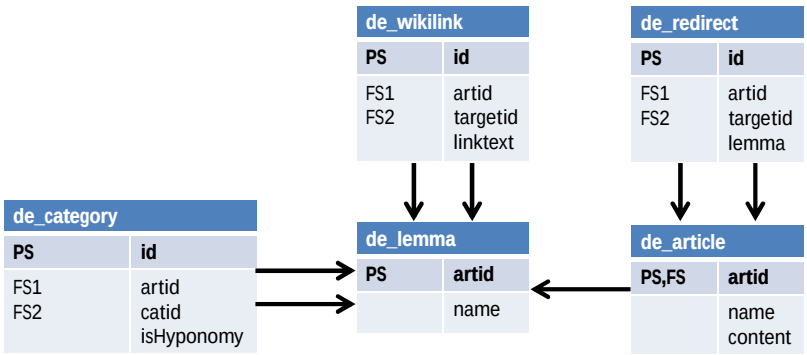

Abbildung 45: Struktur der Datenbanktabellen

4 `http://dumps.wikimedia.org/` - Zugriff am 14.11.2012

6.2 ERWEITERUNG DES DATENMODELLS UND REALISIERUNG VON EMPFEHLUNGEN

In Kapitel 4 wurde das Grundkonzept dieser Arbeit beschrieben, das darin besteht erkannte Hyponomiebeziehungen zwischen Tags dem CROKODIL-Wissensnetz hinzuzufügen und auf Basis dieser hinzugefügten Relationen dem Lernenden Ressourcenempfehlungen anzubieten. Diese beiden Implementierungsaspekte werden nachfolgend dargestellt.

6.2.1 *Erweiterung des Datenmodells*

Die mittels eines Vergleichs mit der Taxonomiedatenbank bestimmten taxonomischen Beziehungen zwischen Tags werden unmittelbar im Datenmodell gespeichert. Ergänzend werden sie dazu benutzt, eine Ähnlichkeit zwischen Ressourcen zu berechnen. Die Ähnlichkeit wird durch die existierende Entfernung in der Taxonomie der die Ressourcen beschreibenden Tags bestimmt. Diese Relation wird später von Expertensuchen benutzt, um neue Empfehlungen zu generieren.[5]

Das bestehende CROKODIL Datenmodell muss daher an zwei Stellen erweitert werden: Die erweiterten und die neuen Klassen sind in Abbildung 46 dargestellt. Es ist eine Assoziationsklasse `is-a` mit einem Attribut `Distanz` neu hinzugekommen. Die Distanz gibt die Entfernung zwischen zwei Tags in der Taxonomie an und wird mit Hilfe einer Breitensuche berechnet (siehe Abschnitt 6.3.2.1). Die Ähnlichkeit zwischen Ressourcen oder allgemeiner Objekten wird mit Hilfe zweier Relationen `quellObjekt` und `aehnlicheObjekte` modelliert.

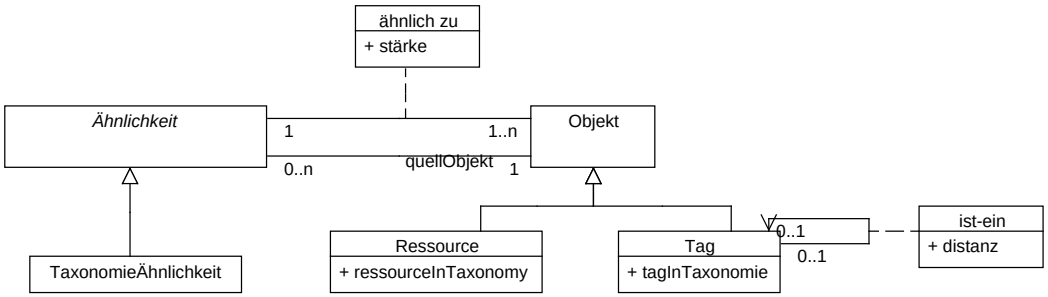

Abbildung 46: Erweiterung des Basismodells der CROKODIL-Plattform

Ähnlich wie bei den *is-a*-Relationen wird auf für die Ähnlichkeitsrelation ein numerisches Attribut, dass die Stärke oder die Ähnlichkeit zwischen den Objekten angibt, benötigt. Die Erstellung der Ähnlichskeitsrelation wird in Detail in Abschnitt 6.3.2.2 erklärt. Zusätzlich ist mittels eines weiteren Attributs `checkTaxonomy` zu speichern, ob die Ähnlichkeit zwischen zwei Objekten, d.h. zwei aus der CROKODIL-Plattform exportierten Tags bereits berechnet wurde.

Bei der Erstellung Ähnlichkeitsrelationen zwischen Ressourcen kann es im schlimmsten Fall passieren, dass Relationen zwischen alle Paare von Ressourcen gezogen werden müssen. In diesem wären die Expertensuchen nicht mehr performant ausführbar. Um dieses Problem umzugehen, wurde die Entscheidung getroffen, die Ähnlichkeits-

5

relation nicht direkt zwischen zwei Objekten zu speichern, sondern Hilfsobjekte zu definieren, die zum einen mit einem Quellobjekt verbunden sind und zum anderen mit allen zu diesem Quellobjekt ähnlichen Ressourcen. Diese Hilfobjekte besitzen zudem einen spezifischen Typ entsprechend des Verfahrens, mittels dessen die Ähnlichkeit berechnet wurde. Konkret wurde eine Taxonomieähnlichkeitsklasse definiert, deren Instanzen sowohl mit einem sogenannten Quellobjekt (Quellobjekt-Relation) als auch mit einem oder mehreren Objekten, die ähnlich zum Quellobjekt sind, über Relationen verbunden sind. An dieser Stelle soll das mit Hilfe eines Beispiels gezeigt werden:

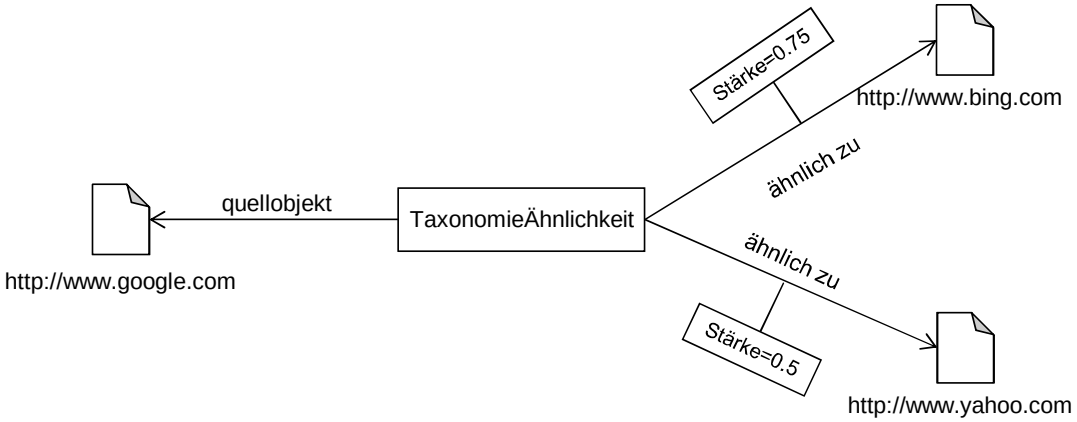

Abbildung 47: Beispiel der Benutzung des Ähnlichkeitsobjekts

Die Ressource „http://www.google.com“ ist das Quellobjekt einer Taxonomieähnlichkeit, die zu zwei ähnlichen Objekten führt. Darüber hinaus lässt sich die Stärke als Attribut der „ähnliches zu“-Relation speichern, sodass die Ähnlichkeit zwischen verschiedenen ähnlichen Objekten verglichen werden kann.

6.2.2 *Generierung von Empfehlungen*

Empfehlungen werden in der CROKODIL-Plattform grundsätzlich mit Hilfe von Expertensuchen generiert. Das gilt bereits für die in Abschnitt 4.1.2.5 beschriebenen Empfehlungen. An dieser Stelle wird erläutert, wie die Ähnlichkeitsrelation zwischen Objekten verwendet wird, um weitere Empfehlungen zu generieren.

Eine Expertensuche (siehe Abb. 48) wird benutzt, um alle Ähnlichkeitsobjekte zu einem gegebenen Objekt zurückzuliefern. Das Objekt beschreibt den „Kontext“ der aktuellen Suche.

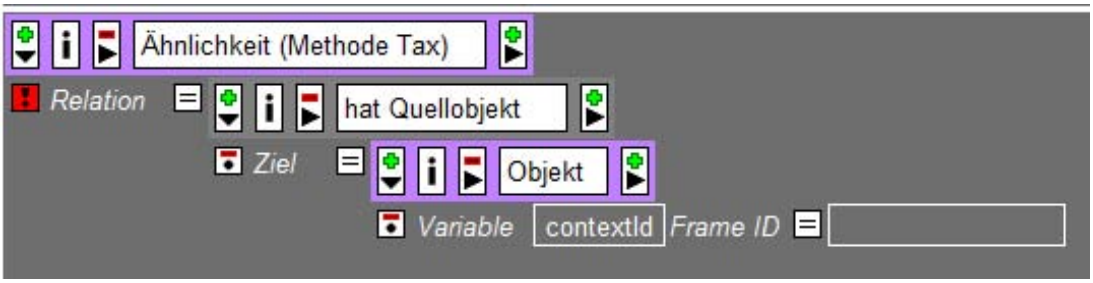

Abbildung 48: Expertensuche nach Objekten der Taxonomieähnlichkeit

Da in der CROKODIL-Plattform verschiedene Algorithmen zur Berechnung von Ähnlichkeiten realisiert sind, können diese Ergebnisse vereinigt, anhand ihrer Stärke (z.B. ist dies abhängig von der Distanz in der Taxonomiedatenbank) gerankt und dem Benutzer angezeigt werden.

Allerdings muss sichergestellt werden, dass Ressourcen in der gleichen Aktivität oder in Ober- und Unteraktivitäten nicht empfohlen werden, da diese keinen Mehrwert für den Benutzer bringen. Zu diesem Zweck wird eine zweite Expertensuche gebildet (siehe Abb. 49) und anschließend die Differenz zwischen der ersten und der zweiten Ergebnismenge gebildet. Das Ergebnis wird dem Benutzer als Empfehlung angezeigt.

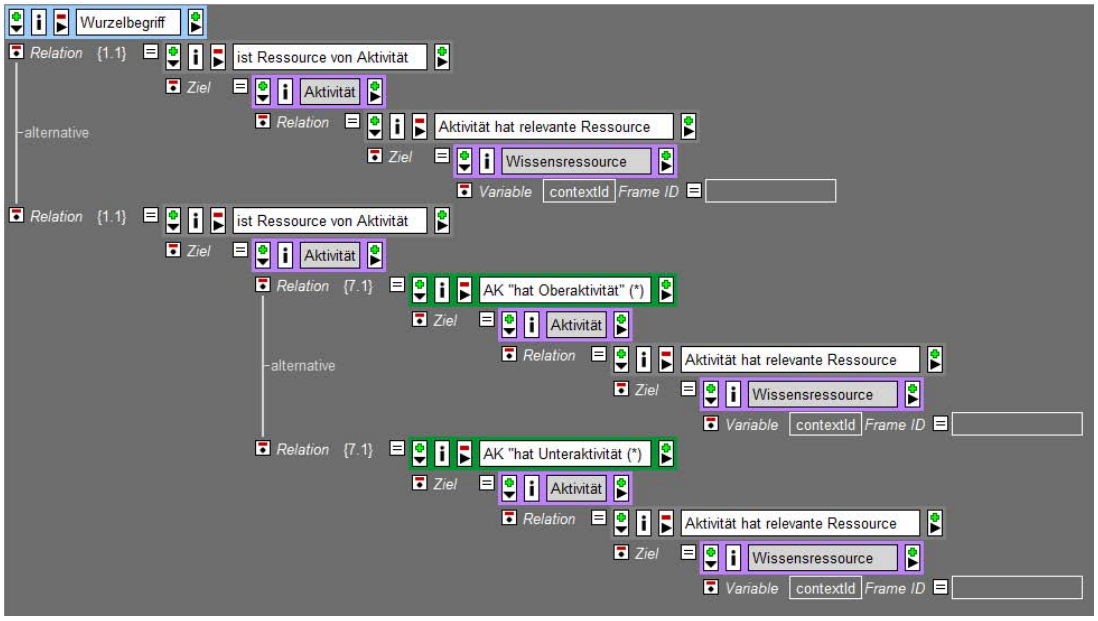

Abbildung 49: Expertensuche nach Ressourcen in der gleichen Aktivität

Die Empfehlungen erscheinen in der CROKODIL-Plattform als sogenannte Kontextboxen. Abbildung 50 zeigt einen CROKODIL-Screenshot zur Ressource „Bing“. Auf der rechten Seite sieht man eine Kontextbox „Empfehlungen“.

In Abbildung 51 wird diese Kontextbox größer dargestellt. Sie zeigt eine Empfehlung auf die ähnliche Wissensressource „Google“.

Diese Empfehlung wird angezeigt, weil die Ressourcen „Bing“ und „Google“ mit den Tags „Suchmaschine“ bzw. „Google“ getaggt. Aus der Taxonomie ließ sich zudem ableiten, dass „Google“ eine „Suchmaschine“ ist. Aus diesen Informationen erzeugt das hier vorgestellte Verfahren ein Ähnlichkeitsobjekt zwischen den Ressourcen „Bing“ und „Google“. Abbildung 52 zeigt den Ausschnitt des semantischen Netzes im Knowledge Builder der Relationen zwischen „Bing“ und „Google“.

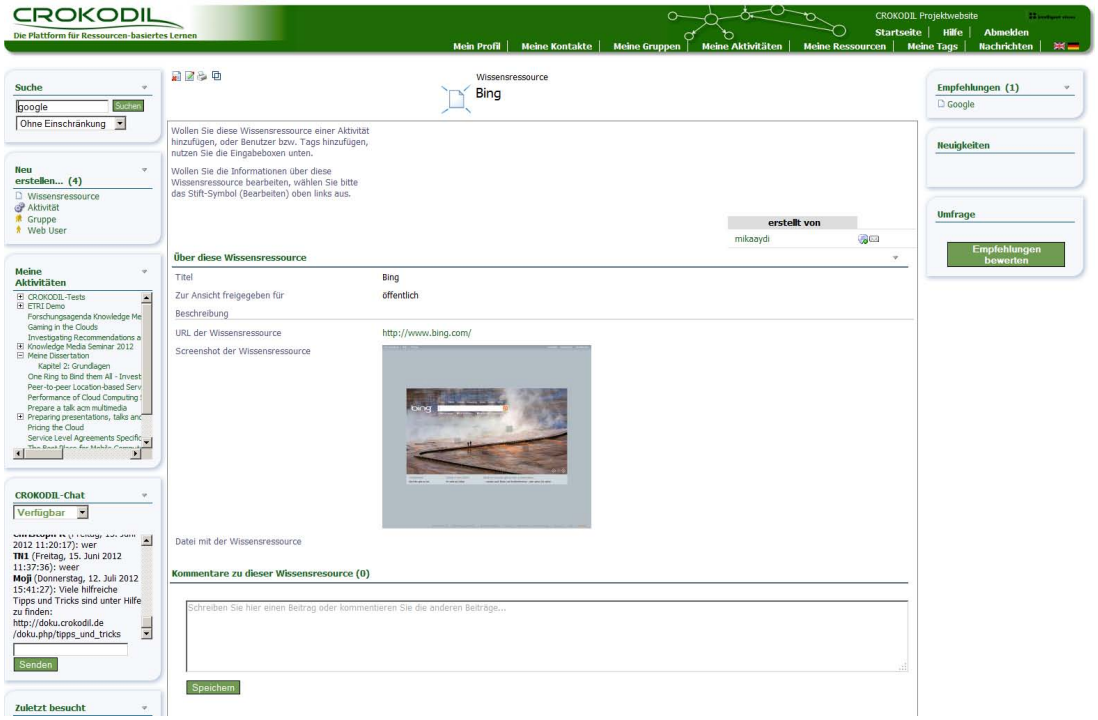

Abbildung 50: Screenshot einer Ressource mit einer Taxonomie-basierten Empfehlung

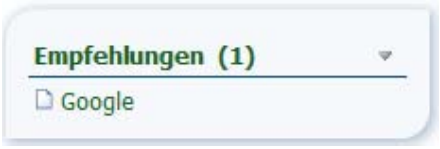

Abbildung 51: Screenshot einer Kontextbox für Empfehlungen

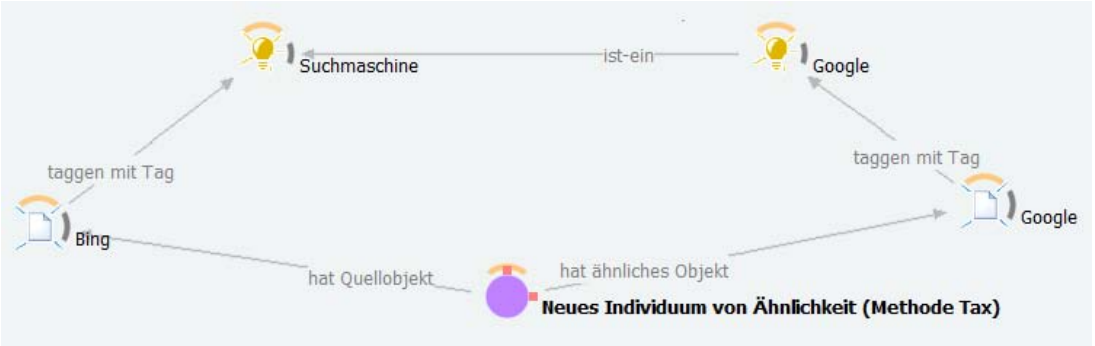

Abbildung 52: Ausschnitt des semantischen Netzes von CROKODIL

6.3 CROKTAXTOOLS

Nachdem im vorhergehenden Abschnitt die Erweiterung des Datenmodells und die Generierung der Empfehlungen mittels der neu definierten Ähnlichkeitsrelationen beschrieben wurde, wird in diesem Abschnitt beschrieben, wie die Relationen generiert und in das CROKODIL Wissensnetz eingefügt werden. Dazu dienen die sogenannten CrokoTaxTools. CrokoTaxTools ist eine Kollektion von Tools und Klassen, die im Rahmen dieser Arbeit implementiert wurden, um folgende Aufgaben durchzuführen:

1. Erkennung von *is-a*-Beziehungen zwischen Tags aus der CROKODIL-Plattform anhand der in der Taxonomiedatenbank gespeicherten und klassifizierten Links des Wikipedia-Kategoriengraphen.
2. Ergänzen von Ähnlichkeitsrelationen zwischen Ressourcen und *is-a*-Relationen zwischen Tags in der CROKODIL-Plattform.
3. Austausch der Daten über Web Services zwischen den verschiedenen Komponenten: K-Infinity und Taxonomiedatenbank.

CrokoTaxTools wurde in der Programmiersprache Java implementiert. Die Berechnung der is-a-Relationen wird einmal am Tag offline durchgeführt und die gefundenen Relationen werden anschließend im semantischen Netz gespeichert. Wenn CrokoTaxTools gestartet wird, werden alle Tags aus der CROKODIL-Plattform über einen GET-Aufruf an den Web-Service geholt. Die Antwort kommt als XML-Datei, die von CrokoTaxTools geparst und verarbeitet wird. Es wird in der Taxonomiedatenbank nach Hyponymierelationen zwischen den Tags (als Wikipedia-Kategorien) gesucht und falls sie gefunden werden, werden diese Relationen über einen POST-Aufruf an dem Web-Service im Netz gespeichert.

6.3.1 *Architektur von CrokTaxTools*

Abbildung 53 zeigt die innere Architektur von CrokTaxTools, bestehend aus vier Paketen und den jeweiligen Klassen. Das Paket `core` stellt die in den anderen Paketen benutzten Datenstrukturen zur Verfügung. Während das Paket `isa` Klassen zur Erkennung der Hyponymien im semantischen Netz beinhaltet, enthält das Paket `similarity` alle Klassen zur Erzeugung der Ähnlichkeitsrelationen zwischen Objekten. Die Kommunikationsaufgaben und verwandte Funktionen werden in Klassen vom Paket `tools` durchgeführt.

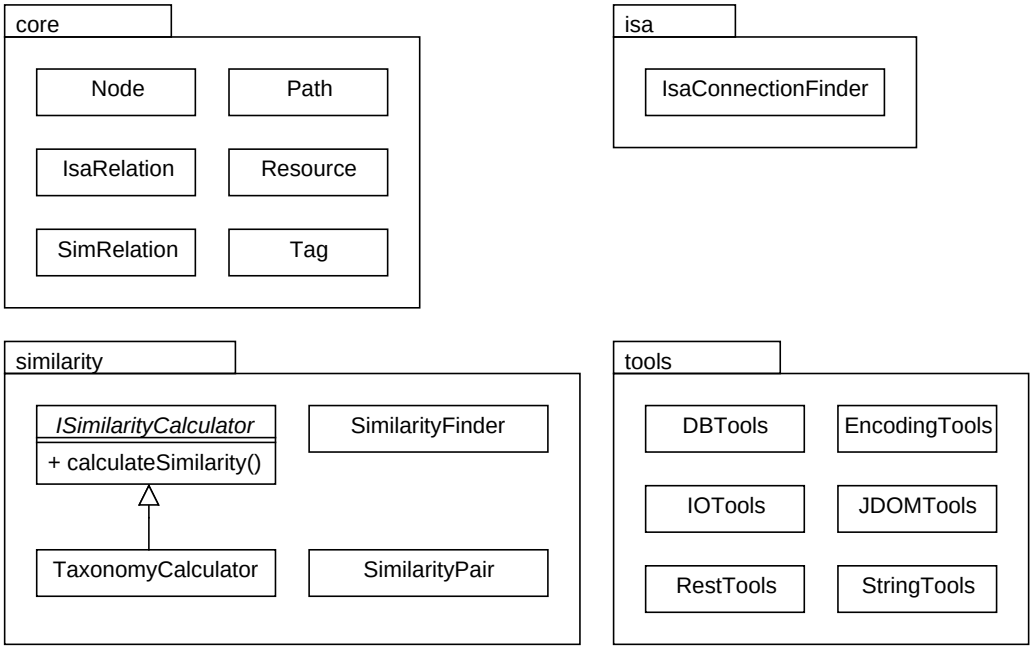

Abbildung 53: Innere Architektur von CrokTaxTools

6.3.2 *Funktionsweise*

Die Funktionen und Funktionsweise von CrokTaxTools sollen in diesem Abschnitt anhand eines Beispiels (vgl. Abb. 54) gezeigt werden. Es besteht aus drei Ressourcen, die jeweils mit einem Tag verschlagwortet wurden.

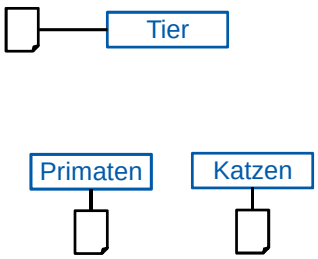

Abbildung 54: Drei Ressourcen mit jeweils einem Tag

6.3.2.1 *Schritt 1: Ziehen von Hyponymierelationen*

Das Ziehen von Hyponymierelationen wird in Algorithmus 6.3.1 dargestellt: Zuerst holt sich das Verfahren alle Tags per REST-Service vom semantischen Netz. Anschließend wird paarweise geprüft, ob in der Taxonomiedatenbank eine *is-a*-Relation zwischen den Tags entsprechender Wikipedia-Kategorien besteht. Wenn einem oder beiden Tags keine Kategorie in der Taxonomiedatenbank entspricht, besteht keine Relation. Mit Hilfe einer Breitensuche kann der kürzeste Weg zwischen beiden Tags bzw. den Kategorien berechnet werden. Dies geschieht abhängig von der Richtung der Relation. Die Information, dass es einen Weg gibt, d.h. eine *is-a*-Relation, wird

dann zusammen mit der Länge des Weges in einer *is-a*-Relation im semantischen Netz gespeichert.

Algorithmus 6.3.1 Pseudocode für die Ziehung von *is-a*-Relationen

Eingabe: Semantisches Netz N, Taxonomie T

```
1:  Prozedur ISA_RELATION_FINDER(N,T)
2:      tags_list = GET_ALL_TAGS_PER_REST(N)
3:      for all (t_i,t_j) ∈ tags_list mit i < j and i ≠ j  do
4:          if ein Weg w zwischen t_i und t_j in T existiert then
5:              d = CALCULATE_LENGTH_OF_PATH(w)
6:              CREATE_ISA_RELATION_PER_REST(t_i,t_j,d,N)
7:          else if ein Weg zwischen t_j und t_i in T existiert then
8:              d = CALCULATE_LENGTH_OF_PATH(w)
9:              CREATE_ISA_RELATION_PER_REST(t_j,t_i,d,N)
10:         else
11:             PRINT_OUT(„KeineHyponymiezwischendenTags“)
```

Ausgabe: Ein verändertes semantisches Netz N′ mit ggf. zusätzlichen is-a-Relationen

Angewendet auf das oben gezeigt Beispiel ergibt sich folgendes semantische Netz:

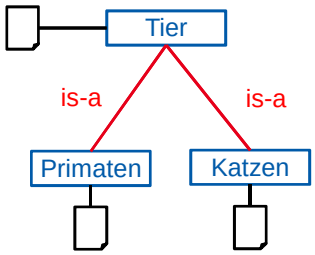

Abbildung 55: Gefundene *is-a*-Relationen

Zwischen den Tags „Tier“ und „Primaten“ und „Tier“ und „Katze“ wird jeweils eine *is-a*-Relation gezogen und zwischen den Tags „Primaten“ und „Katze“ keine. Bei der Berechnung wird jeder Tag mit jedem anderen verglichen ($\frac{n*(n-1)}{2}$ Vergleiche). Tags, die bereits mit den anderen verglichen wurden, werden mit Hilfe des Attributes `tagIntaxonomie` (siehe Abschnitt 6.2.1) markiert, sodass ein Paar aus bereits geprüften Tags in einem späteren Durchlauf nicht mehr getestet wird. Dieses Attribut muss im späteren Verlauf nicht mehr geändert werden.

6.3.2.2 *Schritt 2: Berechnung der Ähnlichkeit und Erstellung von Ähnlichkeitsobjekten*

Das Verfahren holt sich über REST aus dem Wissensnetz alle Tagpaare (t_i,t_j), die über *is-a*-Beziehungen verbunden sind. Zwischen den mittels der Tags verschlagworteten Ressourcen r_i und r_j wird dann ein Ähnlichkeitsobjekt mit Relationen `quellObjekt` zu r_i und `aehnlicheObjekte` zu r_j erzeugt. Die Ähnlichkeitsstärke wird durch die Formel $similarity(r_i,r_j) = 1 - \left\lfloor \frac{d}{hoehe\_taxonomie} \right\rfloor$ berechnet, wobei d die Länge des Pfades zwischen t_i und t_j und $hoehe\_taxonomie$ die Höhe der Taxonomie[6] darstellt. Die Berechnung wird in Algorithmus 6.3.2 dargestellt. Für den besonderen

6 Längster Pfad zwischen Wurzel und Blätter

Fall, dass Ressourcen mehrere Tags gemeinsam haben, wird der Durchschnitt aller Ähnlichkeitswerte berechnet (Zeilen 9 - 11).

Algorithmus 6.3.2 Pseudocode für die Erstellung von Ähnlichkeitsobjekten

Eingabe: Semantisches Netz N, Taxonomie T

1: **Prozedur** CREATE_SIMILARITY_RELATIONS_PER_REST(N,T)
2: isa_list = GET_ALL_ISA_RELATIONS_PER_REST(N)
3: **for all** $(t_i,t_j) \in isa\_list$ **do**
4: **for all** r_i of t_i **do**
5: **for all** r_j of t_j **do**
6: **if** $similarity(r_i,r_j)$ noch nicht existiert **then**
7: $similarity(r_i,r_j) = 1 - \left\lfloor \frac{d}{hoehe\_taxonomy} \right\rfloor$
8: **else**
9: $similarity_{old}(r_i,r_j) = similarity(r_i,r_j)$
10: $similarity_{new}(r_i,r_j) = 1 - \left\lfloor \frac{d}{hoehe\_taxonomy} \right\rfloor$
11: $similarity(r_i,r_j) = \frac{similarity_{old}(r_i,r_j)+similarity_{new}(r_i,r_j)}{2}$
12: CREATE_SIMILARITY_RELATIONS_PER_REST(r_i,r_j,$similarity$,N)

Ausgabe: Ein verändertes semantisches Netz N′ mit ggf. zusätzlichen Ähnlichstsobjekten und -relationen

Für das obige Anwendungsbeispiel wird angenommen, dass die Pfade zwischen „Tier“ und „Primaten“ und „Tier“ und „Katze“ die Länge 3 bzw. die Länge 4 haben, so dass die entsprechende Länge berechnet werden kann. In Abb. 56 wird das Ergebnis der Berechnung dargestellt. Die roten Kästen stellen die Ähnlichkeitsobjekte und die Kanten die Ähnlichkeitsrelationen mit der jeweiligen Ähnlichkeitsstärke dar.

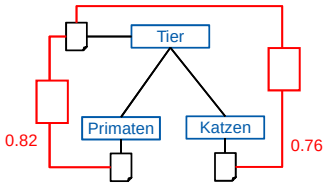

Abbildung 56: Berechnete Ähnlichskeitsrelationen und -objekte

6.4 ZUSAMMENFASSUNG

In diesem Kapitel wurden die im Rahmen dieser Arbeit implementierten Erweiterungen der CROKODIL-Plattform vorgestellt. Mittels dieser Erweiterungen werden neue gewichtete Ähnlichkeitsbeziehungen zwischen Ressourcen in CROKODIL sowie is-a Relationen zwischen Tags berechnet und im CROKODIL Wissensnetz gespeichert. Die Ähnlichkeitsrelationen werden von einer neu definierten Expertensuche in CROKODIL verwendet um zusätzliche Ressourcenempfehlungen für die Lernenden zu bestimmen. Das folgende Kapitel beschreibt die Evaluation der hier vorgestellten Erweiterungen.

7 EVALUATION DER NUTZUNG DER TAXONOMIE IM ANWENDUNGSSZENARIO

»Miss alles, was sich messen lässt, und mach alles messbar, was sich nicht messen lässt.«

— Galileo Galilei

Im Rahmen dieser Arbeit wurden offene Herausforderungen beim kollaborativen Ressourcen-basierten Lernen in Online Communities mit Hilfe eines Taxonomie-basierten Empfehlungssystems adressiert. Das Anwendungsszenario (die CROKODIL-Plattform) und die offenen Herausforderungen wurden in Kapitel 4 vorgestellt und analysiert. Mittels der Umsetzung des in Kapitel 4 vorgestellten Konzepts sollen automatisiert Relationen zwischen zur Verschlagwortung verwendeten Tags im semantischen Netz ergänzt werden, so dass Lernende von den von der Community gesammelten Ressourcen mehr profitieren, generellere und spezifische Ressourcen erkannt werden können und der Suchraum für potentiell relevante Lernressourcen erweitert werden. Die in diesem Kapitel vorgestellte Evaluation zeigt, dass Informationen einer Taxonomie benutzt werden können, um semantische Netze anzureichen, so dass Empfehlungssysteme davon profitieren können. Dieses Kapitel ist wie folgt gegliedert: Der erste Abschnitt befasst sich Verfahren zur Evaluation von Empfehlungssystemen. Insbesondere werden die Grundlagen vorgestellt, um die im Rahmen dieser Arbeit verwendete Evaluationsmethodik zu begründen. Im Abschnitt 7.2.1 werden die in der Evaluation benutzten Korpora vorgestellt und deren Erstellungsprozesse erläutert. Für die Evaluation wurden zusätzliche Algorithmen und Tools benutzt. Diese werden in Abschnitt 7.2.2 beschrieben. Die Ergebnisse werden in Abschnitt 7.3 vorgestellt und diskutiert. Abschließend erfolgt eine Zusammenfassung, in der auf die im Kapitel 4 genannten Ziele eingegangen wird.

7.1 GRUNDLAGEN DER EVALUATION VON EMPFEHLUNGSSYSTEMEN

Grundsätzlich lassen sich Empfehlungssysteme entweder mit Hilfe von historischen Daten oder mit Hilfe von Benutzerbefragungen evaluieren [131]. Im Falle der Verwendung historischer Daten wird die Güte des Empfehlungssystems durch Offline durchgeführte Experimente bestimmt. Bei diesen Experimenten lassen sich die Maße Precision, Recall und F-Maß (vgl. Abschnitt 2.2.3) für das Empfehlungssystem bestimmen. In Benutzerevaluationen werden Benutzer befragt, ob sie einzelne Empfehlungen oder das Empfehlungssystem insgesamt gut finden.

7.1.1 *Evaluation mit historischen Daten*

Die Evaluation mittels historischer Daten ist heutzutage die meistbenutzte Technik zur Evaluation von Empfehlungssystemen. Eine Studie in [65] hat alle im Zeitraum

2004 - 2007 im renommierten Journal ACM Transactions on Information Systems[1] (ACM TOIS) publizierten Forschungsartikel über Empfehlungssysteme analysiert. Drei Viertel aller Artikel wurden mit Hilfe von historischen Daten evaluiert. Die Evaluation anhand historischer Daten gilt als vergleichsweise einfach durchführbar und beliebig wiederholbar. Ein Algorithmus kann mit beliebig vielen Parametrisierungen getestet werden. Darüber hinaus lässt sich auf diese Weise die Qualität von zwei Empfehlungssystemen unter den genau gleichen Bedingungen vergleichen.

Bei der Evaluation mit historischen Daten werden die Evaluationsmetriken und -maße aus dem Information Retrieval (vgl. Abschnitte 2.2.3) benutzt, um sicherzustellen, dass die Messungen verlässlich sind und nicht durch Ausreißer verfälscht werden. Dafür baut die Evaluation von Empfehlungssystemen grundsätzlich auf die im Abschnitt 2.2.4 vorgestellte k-fache stratifizierte Kreuzvalidierung oder Varianten dieser [134] auf. Robuste und stabile Aussagen werden durch die zufällige Aufteilung der Proben und die Tatsache, dass die Experimente k-mal durchgeführt werden, erreicht. In diesem Vorgehen liegt auch der Nachteil von Evaluationen mit historischen Daten: Die zur Evaluation verwendeten Daten stammen aus Anwendungen, die ohne Empfehlungssystem oder mit Hilfe anderer Empfehlungssysteme erhoben wurden. Das bedeutet, dass ein neues Empfehlungssystem hohe Precision- und Recall-Werte liefert, wenn es dieselben Items wie die alten (bzw. keine) Empfehlungssysteme empfehlen kann. Dies führt aber dazu, dass Empfehlungssysteme, die neue (unbekannte) Items empfehlen, nicht mit Hilfe historischer Daten evaluiert werden können. Empfehlungssysteme, die auf die Interaktion mit den Benutzern basieren (wie z.B. [23, 26, 44]), lassen sich ebenfalls nicht mit historischen Daten evaluieren. Diese Empfehlungssysteme erwarten Angaben von Benutzer (oft in Form von Fragen), um Empfehlungen generieren zu können.

7.1.2 *Benutzerevaluationen*

Benutzerevaluationen lassen sich in Online-Evaluationen und Benutzerstudien einteilen [131]. Beide haben gemeinsam, dass die Qualität der Empfehlungen durch die Benutzer explizit oder implizit angegeben oder bestimmt wird. Online-Evaluationen bezeichnen Studien, die im laufenden Betrieb einer Anwendung gemacht werden. Das Ziel ist es, aus dem Verhalten des Benutzers Aussagen abzuleiten. Zum Beispiel, wenn die Benutzer eines Systems einen bestimmten Typ von Empfehlungen mehr anklicken als einen anderen, dann lässt sich schlussfolgern, dass dieses Empfehlungssystem den anderen überlegen ist. Online-Evaluationen mit realen Benutzern gelten als der strengste Beweis der Qualität eines Empfehlungssystems [131]. Allerdings sind die Ergebnisse einer Online-Evaluation nur aussagekräftig, wenn sich viele Benutzer daran beteiligen. Darüber hinaus hängen die Ergebnisse von vielen Faktoren, wie dem Benutzerverhalten, dem Benutzer-Kontext, dem Interface und der Usability der Plattform, ab. Zusammenfassend lässt sich sagen, dass Online-Evaluationen eher benutzt werden, wenn die Wirksamkeit von Empfehlungssystemen im Langzeiteinsatz evaluiert werden soll [131].

Eine Benutzerstudie wird typischerweise durchgeführt, indem eine Menge von Probanden gebeten wird, bestimmte Aufgaben mit Hilfe eines Empfehlungssystems zu erledigen. Während die Probanden die verschiedenen Aufgaben erledigen, wird

1 `http://tois.acm.org/` - Zugriff am 14.11.2012

ihr Verhalten beobachtet und dokumentiert. Die Aufgaben hängen vom Typ der zu empfehlenden Items ab: Es kann sich z.B. um die Auswahl eines Internetproviders anhand von Empfehlungen [43] oder die Suche nach Produkten in einem Online-Katalog [119], handeln. Die Ergebnisse (z.B. die Anzahl der korrekt bearbeiteten Aufgaben oder wie schnell eine Aufgabe gelöst werden konnte) werden dann numerisch ausgewertet und ggf. mit den Ergebnissen mit einer Baseline oder einem anderen Empfehlungssystem verglichen. Benutzerstudien haben den Vorteil, dass sich das Verhalten der Benutzer gut auswerten und vergleichen lässt, haben aber auch einige Nachteile. Benutzerstudien sind oft sehr kostspielig, da die Probanden zuerst angeworben und häufig bezahlt werden müssen. Aus diesem Grund können die Szenarien nur wenige Male durchgespielt werden, was die Verlässlichkeit der Ergebnisse vermindert [131]. Darüber hinaus ist die Auswahl der Probanden von höchster Wichtigkeit, da sie die Benutzer des realen Systems repräsentieren sollen. Ansonsten wären die Ergebnisse nicht vertrauenswürdig.

7.1.3 *Fazit*

In diesem Abschnitt wurden die beiden grundsätzlichen Methoden zur Evaluation von Empfehlungssystemen vorgestellt und deren Vor- und Nachteile diskutiert. Die Evaluation mit Hilfe von historischen Daten überprüft, ob bereits gefundene Items in einem Datensatz gefunden werden. Das im Rahmen dieser Arbeit entwickelte Konzept hat zum Ziel, generelle und spezifische Ressourcen zu empfehlen. Aufgrund der Tatsache, dass nur eine kleinere Menge an Ressourcen zu empfehlen ist, werden die Precision- und Recallwerte eher niedrig sein. Dieses Ergebnis ist unkritisch, weil ein strukturbasiertes Verfahren nicht alleine das Ressourcen-basierte Lernen unterstützen kann. Es sind ergänzend weitere Empfehlungssysteme nötig, wie z.B. Empfehlungen auf Basis der Ähnlichkeit von Ressourcen [146] oder aktivitätsbasierte Empfehlungen [9]. Eine Benutzerevaluation würde unter demselben Problem leiden: Entweder würden die Benutzer nicht optimal in ihrem Lernprozess unterstützt, wenn nur das strukturbasierte Verfahren benutzt wird, oder wenn mehrere Empfehlungssystem im Einsatz sind, dann ist die Messung der Wirkung von einzelnen Empfehlungssystemen in Benutzerstudien grundsätzlich schwer nachweisbar [131].

Aus diesem Grund konzentriert sich die in diesem Kapitel beschriebene Evaluation darauf nachzuweisen, dass Hyponymien benutzt werden können, um semantische Netze anzureichen, so dass Empfehlungssysteme von der umfangreichreicheren Struktur profitieren können. Die genutzte Evaluationsmethodik wird im nächsten Abschnitt vorgestellt.

7.2 ZIELE UND EVALUATIONSMETHODIK

Die im Rahmen dieser Arbeit adressierten Forschungsfragen wurden in Kapitel 4 wie folgt formuliert:

- Wie können alle Lernenden trotz der Verwendung unterschiedliche Begriffe von den in der Community gespeicherten Informationen profitieren, indem sie auf Ressourcen von anderen Lernenden aufmerksam gemacht werden.

- Wie kann die Menge der potentiell relevanten Lernressourcen in den Suchergebnissen erweitert werden, wenn eine Suche keine oder wenige Treffer liefert, indem dem Lernenden zum Beispiel nicht nur Treffer angezeigt werden, die seinen Suchstring enthalten.

- Wie lassen sich hierarchische Strukturen zwischen Themen erkennen, um generelle von spezifischen Ressourcen zu unterscheiden und diese Information den Lernenden zu geben

Zum Nachweis der Nützlichkeit der Taxonomie zur Erweiterung des Wissensnetzes und der darauf basierenden Empfehlungen wurde in der vorliegenden Arbeit eine mehrstufige Evaluationsmethodik angewendet. Im ersten Schritt werden die zuvor genannten Ziele in messbare Evaluationsziele übersetzt, so dass eine Verbesserung tatsächlich festegestellt werden kann. Lernende können von den in der Community gespeicherten Ressourcen profitieren, wenn sie strukturbasierte Empfehlungen bekommen. In der vorgestellten Implementierung werden Empfehlungen mit Hilfe von strukturbasierten Expertensuchen (siehe Abschnitt 6.2.2) bestimmt. Strukturbasierte Verfahren beruhen auf einer möglichst dichten Struktur des benutzten Graphen [69]. Daher wurden folgende Evaluationsziele definiert:

1. Die Evaluation soll zeigen, dass *is-a*-Relationen tatsächlich die Struktur des vorliegenden Wissensnetzes verdichten (durch mehr Kanten) und auf diese Weise mehr strukturbasierte Empfehlungen bestimmt werden können. Dazu wird die Dichte des Netzes gemessen.

2. Die Evaluation soll zudem zeigen, dass strukturbasierte Empfehlungssysteme von den zusätzlichen *is-a*-Relationen im semantischen Netz profitieren können.

Die Dichte von Graphen ist ein in der Forschung zu strukturbasierten Empfehlungssystemen in verwandten Arbeiten [9, 60, 135] anerkanntes Maß. Wenn strukturbasierte Empfehlungssysteme evaluiert werden, erfasst die Evaluation nicht den gesamten Anwendungsdaten, sondern nur einem sogenannten p-Core von Level k ([15]). Ein p-Core von Level k stellt einen Untergraphen dar, bei dem jeder Knoten mit mindesten k anderen in Verbindung steht. Somit wird eine hohe Dichte im Evaluationskorpus garantiert und es ist sichergestellt, dass falsche Empfehlungen nicht aus der Struktur des Graphen resultieren. In einer realen Anwendung, wie z.B. der CROKODIL-Plattform, ist eine hohe Dichte sehr unwahrscheinlich (siehe 7.2.1).

Die Evaluation erfolgt mittels verschiedener CROKODIL-Korpora: Ein Korpus aus der öffentlichen CROKODIL-Plattform und jeweils zwei weiteren Korpora aus den CROKODIL-Anwendungsszenarien [8]. Auf diese Weise soll sichergestellt werden, dass verschiedene inhaltliche Themenbereiche abgedeckt werden. So ist sichergestellt, dass die Ergebnisse unabhängig von der Inhaltsdomäne und von den angewendeten Szenarien sind. Für jeden Korpus wird gemessen, inwieweit sich der Zusammenhang des semantischen Netzes nach dem Hinzufügen von *is-a*-Relationen verändert. Zum Nachweis des zweiten Evaluationsziels wird gezeigt, dass ein State-of-the-Art strukturbasiertes Empfehlungssystem mittels der ergänzten *is-a*-Relationen verbessert werden kann.

Zur Bewertung werden Evaluationsmaße aus dem Information Retrieval und Verfahren der Evaluation mit historischen Daten benutzt (vgl. Abschnitte 2.2 und 7.1.1). Mittels dieses Vorgehens lassen sich anhand der größeren Anzahl von Experimenten

validere Aussagen zur Nützlichkeit des Ansatzes machen. In den folgenden Unterabschnitten werden die Extraktion der Korpora und weitere im Rahmen der Evaluation benutzte Hilfsverfahren beschrieben.

7.2.1 *Auswahl und Erzeugung der Korpora*

Die CROKODIL-Plattform bildet, wie in Kapitel 4 erläutert, das Szenario „Ressourcenbasiertes Lernen in Online-Communities“ ab. Aus diesem Grund wurde das Konzept in der CROKODIL-Plattform integriert und evaluiert.

7.2.1.1 *CROKODIL-Korpora*

Insgesamt wurden drei verschiedene Korpora aus der CROKODIL-Plattform extrahiert und zum Zwecke der Evaluation aufbereitet:

- CROK_p steht für den aus der öffentlichen CROKODIL-Instanz extrahierten Korpus. In dieser Instanz können registrierte Benutzer beliebige Ressourcen speichern und verwalten.
- CROK_1 entspricht dem in den Szenarien 3 und 4 in [8] ermittelten Korpus. Bei den Nutzern handelt es sich um Lernende, die in duale Studiengänge im Bereich der „Business Administration“ und der „Elektro- und Informationstechnik“ eingeschrieben sind.
- CROK_2 bezeichnet den Korpus, der in den Szenarien 1 und 2 aus dem gleichen Paper entstanden ist. In den Szenarien handelt es sich um Umschulungen im Bereich Informationstechnologie mit Lernphasen von maximal einem Tag.

In der CROKODIL-Plattform sind die Daten in einem semantischen Netz gespeichert. Um zur Erreichung des Evaluationsziels 2 die CROKODIL-Korpora auf bestehende Empfehlungssystem anzuwenden, wird das semantische Netz in eine Folksonomie überführt. Die Überführung des semantischen Netzes der CROKODIL-Plattform in eine Folksonomie erfolgte bereits in anderen Arbeiten [9, 134]. In Abschnitt 2.3.8 wurde eine Folksonomie formell als ein 4-Tupel definiert: $F = \{U,T,R,Y\}$, wobei U die endliche Menge der Benutzer, T die endliche Menge der Tags, R die endliche Menge der Ressourcen in der Folksonomie darstellt. Y ist eine ternäre Relation $Y \in U \times R \times T$, die die Tag-Zuweisungen von Benutzern an Ressourcen repräsentiert. Die Menge P ist dabei die Menge der Posts.

Um Empfehlungssalgorithmen, die auf Basis von Folksonomien arbeiten, bei der Evaluation im Rahmen dieser Arbeit verwenden zu können, ergibt sich die Einschränkung, dass Kanten die Form $(u,t,r) \in Y$ (für einen Benutzer u, ein Tag t und eine Ressource r) haben müssen und somit Kanten zwischen Tags nicht erlaubt sind. Daher wird die folgende Prozedur 7.2.1 benutzt, um in CROKODIL existierende Kanten zwischen Tags als ternäre Relation innerhalb der Korpora der CROKODIL-Plattform abzubilden.

Algorithmus 7.2.1 Prozedur zur Abbildung von Hyponymien als ternäre Relationen

Eingabe: Liste von Paaren aus Tags, die durch eine Hyponymierelation verbunden sind $L = \{(t_1,t_2),...,(t_n,t_m)\}$

Prozedur CONVERTION_TO_FOLKSONOMY_RELATION(L)
  $O = \{\}$ ▷ Ausgabeliste
  for all $(t_i,t_j) \in L$ **do**
    $users_i$ = GET_USERS_USING_TAG(t_i)
    $resources_i$ = GET_USERS_USING_TAG(t_i)
    for all users $u_i \in users_i$ **do**
      for all resources $r_i \in resources_i$ **do**
        e = CREATE_FOLKSONOMY_RELATION(u_i,t_2,r_i)
        $L = L \cup e$
    $users_j$ = GET_USERS_USING_TAG(t_j)
    $resources_j$ = GET_USERS_USING_TAG(t_j)
    for all users $u_j \in users_j$ **do**
      for all resources $r_j \in resources_j$ **do**
        e = CREATE_FOLKSONOMY_RELATION(u_j,t_2,r_j)
        $L = L \cup e$

Ausgabe: Eine Liste mit ternäre Relationen O

Somit wird sichergestellt, dass eine Hyponymie-Relation zwischen den Tags t_1 und t_2 als ternäre Relationen abgebildet wird:

- (u_j,t_1,r_j), wobei u_j und r_j alle Ressourcen bzw. Benutzer, die mit t_2 verbunden sind
- (u_i,t_2,r_i), wobei u_i und r_i alle Ressourcen bzw. Benutzer, die mit t_1 verbunden sind

Auf diese Weise wird das Gewicht einer Hyponymie auf alle Ressourcen und Benutzer übertragen, die mit diesen Tags verbunden sind.

Tabelle 18 zeigt den Umfang der zur Evaluation verwendeten Korpora. |S| (engl. subgraphs) gibt an, aus wieviel nicht verbundenen Teilgraphen die Korpora bestehen.

Tabelle 18: Eigenschaften der benutzten Korpora

Datensatz	|U|	|T|	|R|	|Y|	|P|	Datum	|S|
$CROK_p$	22	300	203	2877	986	01.10.2012	3
$CROK_1$	24	93	113	1701	722	01.10.2012	1
$CROK_2$	29	78	53	439	238	01.10.2012	2

7.2.2 *Verwendete Algorithmen und Tools*

Das zweite zuvor genannte Evaluationsziel besteht darin, den Einfluss von zusätzlichen Hyponymierelationen auf strukturbasierte Empfehlungssysteme zu bestimmen. Dazu soll ein State-of-the-Art für strukturbasierte Empfehlungen Verfahren benötigt.

In [134] gibt Rodenhausen einen Überblick über existierende Ansätze. Rodenhausen kommt zu dem Schluss, dass FolkRank sich am besten für die Aufgabe der Ressourcenempfehlung in CROKODIL eignet. FolkRank zeichnet sich u.a. durch Flexibilität bzgl. der Eingaben des Empfehlungssystems und durch stabile Ergebnisse in verschiedenen Szenarien aus. Aus diesen Gründen wird es auch im Rahmen dieser Arbeit zur Evaluation verwendet.

7.2.2.1 *FolkRank*

FolkRank ist ein graphbasiertes Verfahren, das von Hotho et al. in [60] vorgestellt wurde. Es basiert auf dem PageRank-Algorithmus [111], der für das Ranking von Webseiten entwickelt wurde. Dabei wird das Web als ein ungerichteter Graph, bei dem die Webseiten durch Links (Kanten) verbunden sind, betrachtet. Die zugrundeliegende Idee dahinter ist, dass ein Hyperlink von einer Webseite a zu einer Internetseite b auch die Übertragung von Autorität oder Vertrauen von a nach b darstellt. Auf diese Weise werden Webseiten, die viel referenziert werden, höher gerankt als Webseiten, die wenig referenziert werden.

Die Kernidee des FolkRank-Algorithmus ist die Transformation der graphischen Struktur der Folksonomie (siehe Definition2.3.8) zu einem ungerichteten, gewichteten, tripartiten Graphen $G = (V,E)$. Dabei stellen die Ressourcen, Tags und Benutzer der Folksonomie die Knoten des Graphen dar ($V = U \cup T \cup R$). Die Menge der Kanten wird durch $E = \{\{u,t\},\{t,r\},\{u,r\}|(u,t,r) \in Y\}$ gegeben. Das Gewicht w für jede Kante wird durch die Häufigkeiten in der Menge der Kanten vorgegeben. Das Gewicht für eine Kante zwischen einem Benutzer u und einem Tag t wird durch die Anzahl der Ressourcen bestimmt, die u mit t getaggt hat ($w(u,t) = |\{r \in R|(u,t,r) \in Y\}|$). Entsprechend berechnet man $w(t,r)$ und $w(u,r)$ als die Anzahl der Benutzer, die Ressource r mit t getaggt haben bzw. als die Anzahl der Benutzer u, die die Ressource r getaggt haben.

Für die Berechnung von FolkRank wird der Graph als normalisierte Adjazenzmatrix A dargestellt, sodass jede Spalte 1 ergibt. Der Algorithmus startet mit einem beliebigen Vektor w aus nicht-negativen Zahlen. FolkRank iteriert wie folgt:

$$w \leftarrow dAw + (1-d)p, \text{ wobei}$$

p ein Präferenzvektor mit $||w||_1 = ||p||_1$ ist und
$d \in [0,1]$ ein Parameter, durch den der Einfluss von p begrenzt werden kann.

Auf dieser Basis kann FolkRank ein Knoten-spezifischer Ranking (Benutzer-, Ressourcen, oder Tags-spezifisch) in Folksonomien wie folgt berechnen:

1. p spezifiziert einen Präferenzknoten des Benutzers.
2. w_0 ist das Ergebnis von FolkRank für $d = 1$.
3. w_1 ist das Ergebnis von FolkRank für $d < 1$.
4. $w = w_1 - w_0$ ist der finale Gewichtungsvektor und somit der FolkRank.

Für Zwecke der Evaluation wird im Rahmen dieser Arbeit der Standard-FolkRank-Algorithmus mit einer Standard-Parametrisierung[2] verwendet. Als Präferenz-Vektor

2 http://dev.nepomuk.semanticdesktop.org/wiki/CommunityManager - Zugriff am 14.11.2012

wird immer der Benutzer, für den die Empfehlungen berechnet werden, mit einem doppelten Gewicht (als alle anderen Knoten) ausgewählt.

7.2.2.2 *FReSET*

FReSET[3] ist ein Evaluationsframework zur Bewertung Folksonomie-basierte Empfehlungssysteme, das im Rahmen dieser Arbeit entwickelt wurden [34]. FReSET standardisiert die Evaluation mit historischen Daten und erlaubt die Vergleichbarkeit verschiedener Empfehlungssysteme. Darüber hinaus können Vorverarbeitungsalgorithmen, sogenannte *Filter*, die die Anpassung von Datensätzen erlauben, definiert werden. Beispielsweise kann das p-Core einer Folksonomie berechnet werden, sodass nur Knoten betrachtet werden, die mit mindestens k anderen in der Folksonomie verbunden sind. Des Weiteren können beliebige Empfehlungssysteme in das Framework als Plugin integriert werden.

FReSET berechnet die Standardmaße des Information Retrieval als Evaluationsmaße für die verwendeten Korpora und Verfahren. Eine graphische Ausgabe (vgl. Abb. 57) der Ergebnisse steht zur Verfügung, um die Ergebnisse für den Benutzer verständlicher zu machen.

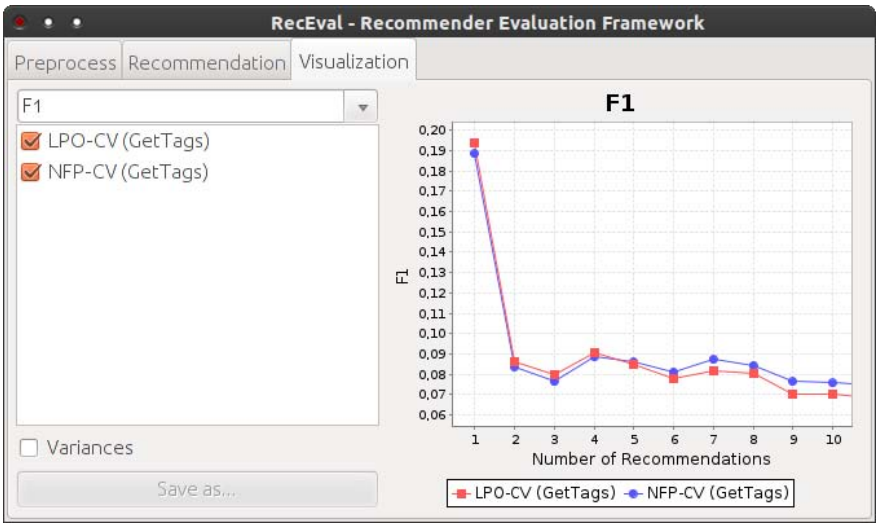

Abbildung 57: FReSET Screenshot des F_1-Graphs

FReSET wurde im Rahmen dieser Evaluation verwendet, um die Qualität von FolkRank bei Verwendung der verschiedenen Korpora zu berechnen, wobei die Korpora mit und ohne ergänzte Hyponymierelationen genutzt werden.

7.3 ERGEBNISSE

In diesem Abschnitt werden die Ergebnisse der Evaluation vorgestellt. Als Erstes wird gezeigt, dass die Folksonomien in den Evaluationskorpora nach Erkennung von *is-a*-Relationen zwischen Tags tatsächlich dichter werden. Anschließend wird auf die Verbesserungen der strukturbasierten Empfehlungen eingegangen.

3 Folsonomy-based REcommender System Evaluation Tool

7.3.1 *Evaluation bzgl. der Dichte des semantischen Netzes*

Zunächst wurde die Dichte der einzelnen Folksonomien gemessen. Die Dichte $dn(G)$ eines Graphen G gibt das Verhältnis der Kantenzahl von G zur Kantenzahl eines vollständigen Graphen[4] [21]. d_{min} und d_{max} stellen den Knoten mit dem jeweils kleinsten bzw. größten Grad dar. Darüber hinaus bezeichnen d_{avg} und d_{med} den Durchschnitt der Grade aus allen Knoten im Graph bzw. den Median der Grade allen Knoten im Graph. Ein Überblick wird in Tabelle 19 gezeigt.

Tabelle 19: Struktur der benutzten Korpora

Datensatz	d_{min}	d_{max}	d_{avg}	d_{med}	dn
$CROK_p$	1	296	16,44	3	12,45
$CROK_1$	1	232	22,19	14	5,51
$CROK_2$	1	45	8,23	3	1,08

Es lässt sich erkennen, dass obwohl der maximale Grad eines Knotens (d_{max}) relativ hoch ist, der Durchschnitt (d_{avg}) und insbesondere der Median (d_{med}) weit davon entfernt liegen. Wie Jäschke et al. in [69] angemerkt haben, leiden viele strukturbasierte Empfehlungssysteme darunter, dass die Graphen nicht dicht sind. Aus diesem Grund betrachten sie nur Korpora, die einen p-Core von mindestens Level 5 haben, um „die Chancen auf gute Empfehlungen zu erhöhen“. Diese Eigenschaft erfüllt keiner der Korpora. Als Folge davon ist es möglich, dass zu manchen Objekten nur entfernte (und somit weniger relevante) Objekte empfohlen werden könnten.

Tabelle 20 zeigt die Eigenschaften der Korpora, nachdem Hyponymierelationen, mittels des in Abschnitt 6.3 vorgestellten Frameworks, erkannt und in der Graphenstruktur eingefügt wurden. In allen Korpora ließen sich Hyponymien finden. Bei den kleineren Korpora wurden weniger Hyponymierelationen gefunden. Durch die Hyponymierelationen wuchsen in allen Korpora sowohl die Anzahl der Kanten als auch die Anzahl der Posts. In den in der Evaluation benutzten Korpora konnte die Anzahl der Untergraphen nicht reduziert werden, da die gefundenen Hyponymierelationen alle im gleichen Untergraphen lagen. Sollten die Tags einer Hyponymie-Relation in verschiedenen Untergraphen liegen, kann auch die Anzahl der Untergraphen reduziert werden.

Tabelle 20: Eigenschaften der benutzten Korpora nach der Erkennung von Hyponymierelationen

Datensatz	$\lvert Y\rvert$	$\lvert Y_{alt}\rvert$	$\lvert P\rvert$	$\lvert P_{alt}\rvert$	$\lvert U\rvert$	$\lvert U_{alt}\rvert$	Hyponymien
$CROK_p$	2951	2877	1042	986	3	3	2
$CROK_1$	1813	1701	786	722	1	1	1
$CROK_2$	443	439	240	238	2	2	1

In $CROK_p$ und $CROK_1$ war der Knoten mit den meisten Kanten Element der erkannten Hyponymierelationen. Dies lässt sich in Tabelle 21 durch die Vergrößerung von d_{max} erkennen. Dies erklärt auch, warum in beiden die Anzahl der Relationen so

4 In vollständigen Graphen ist jeder Knoten mit alle anderen Knoten im Graph direkt verbunden.

gewachsen sind. Bei $CROK_2$ war der Knoten mit den meisten Kanten nicht Element einer Hyponymierelation und somit wuchs die Anzahl der Kanten nur um vier.

Während d_{med} für alle Korpora gleich blieb wuchs, erhöhte sich d_{avg} in jedem Fall.

Tabelle 21: Dichte der benutzten Korpora nach der Erkennung von Hyponymierelationen

Datensatz	d_{max}	$d_{max,alt}$	d_{avg}	$d_{avg,alt}$	d_{med}	$d_{med,alt}$	dn	dn_{alt}
$CROK_p$	301	296	16,86	16,44	3	3	12,77	12,45
$CROK_1$	237	232	23,65	22,19	14	14	6,56	5,51
$CROK_2$	45	45	8,30	8,23	3	3	1,09	1,08

Zusammengefasst lässt sich sagen, dass Hyponymierelationen tatsächlich in der Lage sind, die Struktur der Netze im Anwendungsszenario „Ressourcen-basiertes Lernen in Online Communities“ zu verdichten und somit weitere Empfehlungen zu ermöglichen. In den nächsten Abschnitten soll untersucht werden, ob die Verdichtung der Struktur auch zu besseren Empfehlungen von Ressourcen und Tags in strukturbasierten Empfehlungssystemen führt.

7.3.2 *Empfehlungen anhand eines Empfehlungssystems*

Die Auswirkungen der Ergänzung von Hyponymierelationen in der Folksonomie wurden mit Hilfe von FolkRank bewertet. Dazu wurden zuerst die Gütemaße von FolkRank bei der Verwendung der Korpora ohne Hyponymierelationen berechnet und anschließend für die Korpora mit den erkannten Hyponymierelationen berechnet. Die Ergebnisse wurden wiederum einer zehnfache stratifizierten Kreuzvalidierung unterzogen, in der für jeden Benutzer die Ressourcen bzw. die Tags in 10 Folds aufgeteilt werden und, wie in Abschnitt 2.2.4 erläutert nacheinander, mehrfach berechnet.

In den Tabellen 28, 30, 32 sind die durchschnittlichen Werte der berechneten Maße Precision, Recall, F_1-Maß bei Anwendung von FolkRank auf die Korpora ohne Hyponymien dargestellt. Die Tabellen 22, 24 und 26 zeigen die entsprechenden Ergebnisse für die k-ersten-Empfehlungen. Precision, Recall, F_1-Maß und bei $k = 1$ sind die Gütemaße wenn nur den ersten Treffer einer Empfehlungsliste in die Evaluation einbezogen wird. $k = 10$ bedeutet entsprechend, dass die ersten 10 Treffer für die Evaluation relevant sind. In Evaluation wurde also angenommen, dass höchsten die ersten 10 Empfehlungen für einen Benutzer relevant sind. Höhere Werte von k wären zwar möglich und würden zu besseren Ergebnissen führen, aber laut einer Studie [67] ist es so, dass knapp 70% der Nutzer nur an den ersten 10 Treffern einer Trefferliste interessiert sind.

Tabelle 22: Ergebnisse für $CROK_p$ ohne Hyponymien

k	Precision	Recall	F_1-Maß
1	0,1285	0,0425	0,0556
2	0,1017	0,0414	0,0468
3	0,1224	0,0705	0,067
4	0,1412	0,1274	0,0943
5	0,1548	0,1615	0,1137
6	0,161	0,1881	0,1266
7	0,1679	0,2124	0,1386
8	0,1688	0,2268	0,1454
9	0,1705	0,2386	0,153
10	0,1699	0,2403	0,1574

Tabelle 23: Ergebnisse für $CROK_p$ mit Hyponymien

k	Precision	Recall	F_1-Maß
1	0,1404	0,0449	0,0591
2	0,0938	0,0426	0,0475
3	0,1155	0,0739	0,0691
4	0,142	0,1378	0,102
5	0,1534	0,1621	0,1144
6	0,161	0,1897	0,1277
7	0,1656	0,2083	0,137
8	0,1655	0,2214	0,1426
9	0,16	0,225	0,143
10	0,1651	0,2299	0,1517

Tabelle 24: Ergebnisse für $CROK_1$ ohne Hyponymien

k	Precision	Recall	F_1-Maß
1	0,3234	0,0662	0,1031
2	0,2711	0,1123	0,145
3	0,2604	0,1603	0,1768
4	0,2264	0,1908	0,1793
5	0,199	0,2149	0,1771
6	0,1963	0,2947	0,1963
7	0,1665	0,2113	0,1535
8	0,1694	0,237	0,1645
9	0,1749	0,278	0,1792
10	0,1734	0,2978	0,1836

Tabelle 25: Ergebnisse für $CROK_1$ mit Hyponymie

k	Precision	Recall	F_1-Maß
1	0,3333	0,0561	0,0907
2	0,2917	0,1125	0,1462
3	0,2934	0,1656	0,1871
4	0,2591	0,1887	0,1915
5	0,2354	0,2098	0,1922
6	0,2287	0,2785	0,2083
7	0,2143	0,2498	0,1969
8	0,2138	0,2887	0,2093
9	0,2109	0,3112	0,2151
10	0,2031	0,279	0,2033

Tabelle 26: Ergebnisse für $CROK_2$ ohne Hyponymien

k	Precision	Recall	F_1-Maß
1	0,0398	0,0369	0,0379
2	0,0398	0,049	0,0404
3	0,0398	0,0722	0,0478
4	0,0991	0,1943	0,1277
5	0,1598	0,3941	0,2221
6	0,1411	0,4191	0,206
7	0,1304	0,4337	0,1955
8	0,1187	0,4363	0,1822
9	0,1133	0,4614	0,1777
10	0,1237	0,5522	0,1979

Tabelle 27: Ergebnisse für $CROK_2$ mit Hyponymien

k	Precision	Recall	F_1-Maß
1	0,0618	0,0454	0,05
2	0,0562	0,0601	0,0535
3	0,0487	0,0802	0,0562
4	0,1037	0,1983	0,1322
5	0,1591	0,3839	0,2195
6	0,1462	0,4239	0,2117
7	0,1216	0,4031	0,1817
8	0,1124	0,41	0,1719
9	0,1075	0,4332	0,1679
10	0,122	0,5337	0,1941

Vergleicht man das zusammengefasste F_1-Maß (Tabellen 28 - 33) haben sich die Gesamtergebnisse von FolkRank in den Korpora $CROK_1$ und $CROK_2$ verbessert.

Bei $CROK_p$ haben sich die Durchschnittswerte der Evaluationsmaße bei Berücksichtigung der Hyponymierelationen leicht im dritten Nachkommastellenbereich verschlechtert (siehe Tabelle 29). Allerdings kam es zu einer Verbesserung des F_1-Maßes für Werte $k < 7$, was Precision und Recall in einem Maß vereint, wie man aus dem Vergleich der Tabellen 28 und 29 erkennen kann. Das bedeutet, dass eine Empfehlungsliste mit höchstens 6 Empfehlungen mit Hyponymien mehr relevante Treffer beinhaltet als eine Empfehlungsliste ohne Berücksichtigung der Hyponymiebeziehungen.

Tabelle 28: $CROK_p$: Gesamtergebnisse ohne Hyponymie

Maß	Wert
Precision	0,1486
Recall	0,1546
F_1-Maß	0,1097

Tabelle 29: $CROK_p$: Gesamtergebnisse mit Hyponymie

Maß	Wert
Precision	0,1462
Recall	0,1532
F_1-Maß	0,1092

Die größten Verbesserungen wurden im $CROK_1$-Korpus (Tabellen 30 und 31) beobachtet. Es zeigte sich, dass sich die Precision um 14,1%, der Recall um 4,05% und das F_1-Maß um 14,1 % sich verbesserte (vgl. Tabelle 31). Die detaillierten Ergebnissen (Tabellen 24 und 25) zeigen besonders bei der Precision eine Verbesserung für alle für k betrachteten Werte. Der Recall und das F_1-Maß waren ebenfalls besser, außer für $k = 1$.

Tabelle 30: $CROK_1$: Gesamtergebnisse ohne Hyponymien

Maß	Wert
Precision	0,2192
Recall	0,2027
F_1-Maß	0,1655

Tabelle 31: $CROK_1$: Gesamtergebnisse mit Hyponymien

Maß	Wert
Precision	0,2501
Recall	0,2109
F_1-Maß	0,1831

Der $CROK_2$-Korpus ist durch die geringste Dichte aller im Rahmen dieser Arbeit betrachteten Korpora charakterisiert. Hier lagen die Ergebnisse für die Precision und das F_1-Maß sowohl im detaillierten Vergleich (Tabellen 26 und 27) als auch in den Gesamtergebnissen (Tabellen 32 und 33) bei Einbezug der Hyponymierelationen über den Ergebnissen ohne Einbezug der Hyponymierelationen. Für den Recall waren die Ergebnisse nur für $k < 5$ besser als die Ergebnisse ohne Einbezug der Hyponymierelationen.

Tabelle 32: $CROK_2$: Gesamtergebnisse ohne Hyponymien

Maß	Wert
Precision	0,099
Recall	0,2935
F_1-Maß	0,1403

Tabelle 33: $CROK_2$: Gesamtergebnisse mit Hyponymien

Maß	Wert
Precision	0,1029
Recall	0,2871
F_1-Maß	0,1413

7.4 FAZIT UND DISKUSSION

Im Rahmen dieses Kapitel wurde evaluiert, ob diese Ziele erreicht wurden. Die Evaluation wurde auf drei verschiedenen Korpora durchgeführt. Es wurden zwei konkrete Werte identifiziert, die im Rahmen der Evaluation gemessen wurden (vgl. Abschnitt 7.2). Als Erstes wurde überprüft, ob Hyponymierelationen tatsächlich die Struktur eines semantischen Netzes verändern können. Dies wurde anhand der Messung der Dichte der semantischen Netze mit und ohne Hyponymierelationen gemessen. Die Evaluationsergebnisse zeigten, dass Hyponymie-Beziehungen in allen Korpora gefunden werden konnten und dass sowohl die Dichte der Korpora als auch der einzelnen Knoten erhöht werden konnten. Anschließend wurde geprüft, ob die zusätzlichen Hyponymierelationen im Netz auch eine Auswirkung auf die Güte strukturbasierter Empfehlungsverfahren haben. Zu diesem Zweck wurde die Güte von FolkRank jeweils auf den Korpora mit und ohne Hyponymierelationen berechnet. Auch hier zeigten sich positive Ergebnisse in allen Korpora. Für die Szenarien-spezifischen Korpora $CROK_1$ und $CROK_2$ ließen sich eindeutige Verbesserungen bei den Gesamtergebnissen beobachten. Bei $CROK_p$, also dem Korpus der öffentlichen Plattform, waren die Ergebnisse in den ersten beiden Positionen von Emfpehlungen besser. Es scheint so zu sein, dass es bei spezifischen Szenarien eine höhere Wahrscheinlichkeit gibt, relevante Relationen zu finden. Dies führt dazu, dass die gefundenen Hyponymien potentiell interessante Ressourcen stärker gewichten und diese somit höher

gerankt werden. Der $CROK_p$-Korpus ist heterogener und zeichnet sich durch viele verschiedene Themen aus. Somit haben die gefundenen Hyponymierelationen nur Auswirkungen auf die ersten Treffer der Liste aus. Aus diesem Grund schneiden sie nur bei diesen Werten besser ab. Diese Erkenntnisse wurden mit Hilfe weiterer Experimente mit anderen Parametrisierungen bestätigt (vgl. Anhang A.3.2). Diese Ergebnisse der weiteren Experimenten werden im Anhang A.3.2.2 vorgestellt.

Zusammenfassend zeigt sich, dass erstens mittels Hyponymierelationen das semantische Netz angereichert werden kann und das damit zweitens strukturbasierte Empfehlungssysteme bessere Empfehlungen liefern können. Damit können im Anwendungsszenario des Ressourcen-basierten Lernens Lernender besser auf Ressourcen anderer Lernender hingewiesen werden, was das zentrale Ziel dieser Arbeit war.

8

ZUSAMMENFASSUNG UND AUSBLICK

»Jedes Ziel öffnet den Ausblick auf ein anderes, das auch vergänglich ist.«

— Ralph Waldo Emerson

Dieses Kapitel fasst die Hauptbeiträge dieser Arbeit zusammen und schließt mit einem Ausblick für zukünftige weitere Arbeiten ab.

8.1 ZUSAMMENFASSUNG UND BEITRÄGE DER ARBEIT

Ziel dieser Arbeit war es, das Ressourcen-basierte Lernen innerhalb einer Community von Lernenden zu unterstützen, indem Lernende situationsbezogen auf Wissensressourcen hingewiesen werden, die andere Community-Mitglieder bereits verwendet haben.

Zur Erreichung dieses Ziels wurden das Anwendungsszenario und beispielhaft die das Ressourcen-basierte Lernen unterstützenden Lernumgebung CROKODIL untersucht. Die Untersuchung ergab, dass Benutzer oftmals nicht auf interessante Ressourcen hingewiesen werden können, wenn sie unterschiedliche Terminologien bei der Verschlagwortung von im Lernen genutzten Ressourcen verwenden. Basierend auf dieser Feststellung wurde ein Konzept entwickelt, welche die Lücken in den von den Benutzern verwendeten Terminologien mittels der Verwendung einer Taxonomie schließt. Die Analyse ergab weiterhin, dass das Anwendungsszenario dadurch gekennzeichnet ist, dass die Benutzer aktuelle, teilweise sozio-kulturell spezifische Begriffe in mehreren Sprachen als Schlagworte verwenden. Eine Taxonomie, die diese Schlagworte in Beziehung zueinander setzen will, muss daher dadurch charakterisiert sein, dass sie sehr aktuell ist und im mehreren Sprachen vorliegt. Diese Eigenschaften lassen die Nutzung manuell erzeugter Taxonomien als nicht umsetzbar erscheinen.

Daher wurden in der Arbeit mit TaxWikiHeur.KOM und TaxWikiML.KOM zwei Verfahren konzipiert und implementiert, die weitestgehend sprachunabhängig aus der Online Enzyklopädie Wikipedia Taxonomien generieren, indem sie Links zwischen Wikipedia Kategorien in Hyponymie und Nicht-Hyponymiebeziehungen klassifizieren. Diese Verfahren zeichnen sich dadurch aus, dass sie keine externen, manuell erzeugten Wissensbasen verwenden. Damit besteht keine Notwendigkeit einer manuellen Pflege von Taxonomien für neue Wissensbereiche, die im Anwendungsbereich des selbstgesteuerten Ressourcen-basierten Lernens besonders bedeutsam sind. Das Verfahren TaxWikiML.KOM erweitert das Verfahren TaxWikiHeur.KOM und behebt einige der bei der Evaluation von TaxWikiHeur.KOM erkannten Mängel. Die Evaluation der Verfahren hat insgesamt gezeigt, dass die Güte der Taxonomien sehr gut ist und der Güte automatisch erzeugter Taxonomien (ohne externe Wissensbasen) kaum nachsteht. Die Verwendung der Verfahren erfolgte in fünf Sprachen, so dass der Nachweis der sprachunabhängigen Nutzbarkeit ebenfalls erfolgte.

Das Verfahren TaxWikiML.KOM wurde in der Arbeit weiterhin verwendet, um innerhalb der CROKODIL-Lernumgebung automatisch Beziehungen zwischen von den

Benutzern verwendeten Schlagworten zur Beschreibung der im Lernprozess genutzten Ressourcen zu ergänzen. Das in der Arbeit vorgestellte Konzept zur Anreicherung wurde in Form der CokoTaxTools implementiert und evaluiert. Es konnte zum einen anhand dreier Korpora aus dem Anwendungsfeld der Ressourcen-basierten Lernens nachgewiesen werden, dass die Dichte des Netzes durch das implementierte Konzept größer wird, womit Empfehlungssystemen umfangreichere Informationen zur Generierung von Empfehlungen zur Verfügung stehen, die auch solche Ressourcen anderer Lernender empfehlen können, die mit einer unterschiedlichen Terminologie beschrieben sind. Der positive Einfluss von mittels TaxWikiML.KOM ergänzten Hyponymiebeziehungen zwischen Schlagworten auf die Güte von Empfehlungssystemen wurde in einer weiteren Evaluation anhand des State-of-the-Art Verfahrens FolkRank zusätzlich nachgewiesen. Zur Evaluation wurde das Framework FReSET konzipiert und entwickelt, welches modular und parameterisierbar aufgebaut ist und damit allgemein zur Evaluation von auf Folksonomien basierenden Empfehlungssystemen verwendet werden kann. Mit der Evaluation anhand von FolkRank erfolgte ergänzend der Nachweis, dass das Konzept nicht nur für den konkreten Anwendungsbereich des kollaborativen Ressourcen-basierten Lernens sinnvoll und umsetzbar ist, sondern auch in anderen Anwendungsbereichen nutzbar ist

8.2 AUSBLICK

Im Rahmen dieser Arbeit wurden verschiedene Beiträge in den Bereichen automatische Extraktion von Wissensbasen und Empfehlungssysteme im Ressourcen-basierten Lernen entwickelt und vorgestellt. Diese Beiträge liefern weitere Anknüpfungspunkte für zukünftige Forschungsarbeiten.

Im Bereich der automatischen Informationsextraktion wurden zwei Verfahren zur Erkennung von Hyponymie-Beziehungen in Wikipedia vorgestellt. Insbesondere WikiTaxML.KOM könnte dahin erweitert werden, dass weitere Features entwickelt werden, die z.B. Interwikilinks oder die semantische Ähnlichkeit ausnutzen, um die Genauigkeit des Verfahrens zu verbessern. Die Erkennung von Hyponymie-Beziehungen mittels dieser Verfahren führt, wendet man sie auf die gesamte Wikipedia an, genauso wie in [118], zu einer Großtaxonomie, die nicht zusammenhängend ist. Aus diesem Grund stellt sich die Frage, wie man diese Großtaxonomie zu einer zusammenhängenden Großtaxonomie machen kann. Diese Fragestellung haben Ponzetto und Navigli in [117] gestellt und ein Verfahren für Mapping von Taxonomien mit Hilfe von WordNet, d.h. allein für die englische Sprache, entwickelt. Um die Multilingualität nicht einzuschränken, könnte man ein ähnliches Verfahren mit Hilfe von Wiktionary entwickeln. In Wiktionary werden Hyponymie-Beziehungen zwischen Konzepten von einer großen Community (ähnlich wie Wikipedia) gepflegt.

Die entwickelten Verfahren wurden in dieser Arbeit verwendet, um Empfehlungssysteme für das Ressourcen-basierte Lernen zu verbessern, indem die Benutzer auf Ressourcen anderer Benutzer hingewiesen werden, auch wenn sie andere Terminologien verwenden. Eine Nutzung in anderen Anwendungen, außerhalb des Lernens, in denen Benutzer beliebige Artefakte verschlagworten, erscheint sinnvoll, denn der positive Einfluss des Verfahrens auf die Qualität von Empfehlungen konnte bereits in der Arbeit nachgewiesen werden. Weiterhin denkbar ist bei der Generierung von Empfehlungen im Ressourcen-basierten Lernen nicht nur die Verwendung einer Taxonomie,

sondern auch weitere Community-erzeugter Wissensbasen, wie zum Beispiel die DBPedia oder YAGO. Des Weiteren sollte der Langzeiteinsatz des Taxonomie-basierten Empfehlungssystems im Zusammenspiel mit anderen Empfehlungssystemen bewertet werden.

Darüber hinaus können Hyponymierelationen nicht nur für die Empfehlung von Ressourcen, sondern auch für die Empfehlung von Tags und Benutzer verwendet werden. Die Empfehlung von Tags hat nicht nur einen Beitrag bei der Vereinheitlichung der Terminologie im semantischen Netz [69], sondern führt dazu, dass Hyponymierelationen gleich beim Taggen von Ressourcen gezogen werden können und somit Berechnungsaufwand reduziert werden kann. Mit Hilfe von Benutzer-Empfehlungen kann ein Lernender Benutzer finden, die sich ähnlichen Themen oder spezifischeren Themen beschäftigt haben. Ein Lernender könnte die empfohlenen Benutzer kontaktieren, um das Thema zu diskutieren oder Hilfe bei einem Thema zu bekommen.

Schließlich könnte das FReSET-Tool zur Evaluation von Empfehlungssystemen weiterentwickelt werden. Das Tool wurde bereits in verschiedenen Arbeiten zur Evaluation verwendet, da es einen standardisierten Vergleich von Empfehlungssystemen ermöglicht. FReSET könnte durch weitere Empfehlungssysteme und Filter erweitert werden, sodass es mehr Evaluationsmöglichkeiten gäbe. Darüber hinaus ließen sich neue Evaluationsmaße implementieren und auf ihre Aussagekraft hin überprüfen.

LITERATURVERZEICHNIS

[1] Adafre, S. F. ; Rijke, M. de: Finding Similar Sentences across Multiple Languages in Wikipedia. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, S. 62–69

[2] Adar, E. ; Skinner, M. ; Weld, D. S.: Information Arbitrage across Multi-lingual Wikipedia. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining* ACM, 2009, S. 94–103

[3] Adomavicius, G. ; Tuzhilin, A.: Context-Aware Recommender Systems. In: *Proceedings of the 2008 ACM conference on Recommender systems*, Springer, 2011, S. 217–253

[4] Aehnelt, M. ; Ebert, M. ; Beham, G. ; Lindstaedt, S. ; Paschen, A.: A Socio-Technical Approach towards Supporting Intra-Organizational Collaboration. In: *Times of Convergence. Technologies Across Learning Contexts* 5192 (2008), S. 33–38

[5] Anderson, M. ; Ball, M. ; Boley, H. ; Greene, S. ; Howse, N. ; Lemire, D.: RACOFI: A Rule-Applying Collaborative Filtering System. In: *Proceedings of the International Workshop on Collaboration Agents: Autonomous Agents for Collaborative Environments*, 2003, S. 13

[6] Andreevskaia, A. ; Bergler, S.: Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* Bd. 6, 2006, S. 209–216

[7] Anjorin, M. ; Domínguez García, R. ; Rensing, C.: CROKODIL: a Platform Supporting the Collaborative Management of Web Resources for Learning Purposes. In: *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*. New York, NY, USA : ACM, Jun 2011. – ISBN 978–1–4503–0697–3, S. 361. – Poster

[8] Anjorin, M. ; Rensing, C. ; Bischoff, K. ; Bogner, C. ; Lehmann, L. ; Reger, A. L. ; Faltin, N. ; Steinacker, A. ; Lüdemann, A. ; Domínguez García, R.: CROKODIL - a Platform for Collaborative Resource-Based Learning. In: *Towards Ubiquitious Learning, Proceedings of the 6th European Conference on Technology Enhanced Learning*. Heidelberg : Springer, Sep 2011 (LNCS 6964). – ISBN 9783642239847, S. 29–42

[9] Anjorin, M. ; Rodenhausen, T. ; Domínguez García, R. ; Rensing, C.: Exploiting Semantic Information for Graph-Based Recommendations of Learning Resources. In: *Proceedings of the 7th European Conference on Technology Enhanced Learning*. Heidelberg : Springer, Sep 2012

[10] Arborea, V.: *Wissensportal: Taxonomie*. Webseite, 2012. – Online verfügbar unter `http://www.verein-arborea.de/wissen/taxonomie`; Zugriff am 14.11.2012.

[11] Auer, S. ; Bizer, C. ; Kobilarov, G. ; Lehmann, J. ; Cyganiak, R. ; Ives, Z.: Dbpedia: A nucleus for a web of open data. In: *Proceedings of the 6th International Semantic Web Conference*, Springer, 2007, S. 722–735

[12] Auer, S. ; Lehmann, J.: What have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In: *Proceedings of the 4th European Semantic Web Conference*, Springer, 2007, S. 503–517

[13] Augustin, A. G.: *Erfassung Semantischer Informationen aus Enzyklopädischen Daten*, Technische Universität Graz, Diplomarbeit, 2012

[14] Barritt, C. ; Lewis, D. ; Wieseler, W.: *Cisco Systems Reusable Information Object Strategy*. Webseite, Oktober 1999. – Zugriff am 28.11.2012

[15] Batagelj, V. ; Zaveršnik, M.: *Generalized Cores*. 2002

[16] Bengio, Y. ; Grandvalet, Y.: No Unbiased Estimator of the Variance of K-Fold Cross-Validation. In: *The Journal of Machine Learning Research* 5 (2004), S. 1089–1105

[17] Böhnstedt, D. ; Scholl, P. ; Benz, B. ; Rensing, C. ; Steinmetz, R. ; Schmitz, B.: Einsatz persönlicher Wissensnetze im Ressourcen-basierten Lernen. In: *DeLFI 2008: 6. e-Learning Fachtagung Informatik*. Köllen, Bonn : Lecture Notes in Informatics (LNI), Sep 2008 (P-132). – ISBN 978–3–88579–226–0, S. 113–124

[18] Böhnstedt, D. ; Scholl, P. ; Rensing, C. ; Steinmetz, R.: Collaborative Semantic Tagging of Web Resources on the Basis of Individual Knowledge Networks. In: *Proceedings of First and Seventeenth International Conference on User Modeling, Adaptation, and Personalization* Bd. Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg 2009, Jun 2009. – ISBN 978–3–642–02246–3, S. 379–384

[19] Böhnsteedt, D.: *Semantisches Tagging zur Verwaltung von webbasierten Lernressourcen*, TU Darmstadt, Diss., Juni 2011

[20] Bollacker, K. ; Evans, C. ; Paritosh, P. ; Sturge, T. ; Taylor, J.: Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* ACM, 2008, S. 1247–1250

[21] Borgatti, S. P.: *Graph theory*. Reading List CAMOS, School of Computer Science, Carnegie Mellon University,

[22] Boyle, T.: Design Principles for Authoring Dynamic, Reusable Learning Objects. In: *Australian Journal of Educational Technology* 19 (2003), S. 46–58

[23] Bridge, D. ; Ricci, F.: Supporting Product Selection with Query Editing Recommendations. In: *Proceedings of the First ACM Conference on Recommender systems* ACM, 2007, S. 65–72

[24] Budanitsky, A. ; Hirst, G.: Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. In: *Workshop on WordNet and Other Lexical Resources* Bd. 2, 2001

[25] Burke, R.: Knowledge-Based Recommender Systems. In: *Encyclopedia of Library and Information Systems* 69 (2000), Nr. Supplement 32, S. 175–186

[26] Burke, R. ; Hammond, K. J. ; Young, B. C.: The FindMe Approach to Assisted Browsing. In: *IEEE Expert* 12 (1997), Nr. 4, S. 32–40

[27] *Kapitel* 1. In: Candillier, L. ; Jack, K. ; Fessant, F. ; Meyer, F.: *State-of-the-Art Recommender Systems*. IGI Global, 2009, S. 1–22

[28] Chatti, M.A. ; Dyckhoff, A.L. ; Schroeder, U. ; Thüs, H.: Forschungsfeld Learning Analytics. In: *i-com – Zeitschrift für interaktive und kooperative Medien* 11 (2012), Nr. 1, S. 22–25

[29] Chen, C.M. ; Lee, H. M. ; Chen, Y. H.: Personalized E-Learning System using Item Response Theory. In: *Computers & Education* 44 (2005), Nr. 3, S. 237–255

[30] Chernov, S. ; Iofciu, T. ; Nejdl, W. ; Zhou, X.: Extracting Semantic Relationships between Wikipedia Categories. In: *Proceedings of the First International Workshop on Semantic Wikis* Citeseer, 2006

[31] Cimiano, P. ; Pivk, A. ; Schmidt-Thieme, L. ; Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: *Ontology Learning from Text: Methods, Evaluation and Applications* 123 (2005), S. 59–73

[32] Cormen, T.H. ; Leiserson, C. E. ; Rivest, R. ; Stein, C.: *Algorithmen-Eine Einführung*. Oldenbourg Wissenschaftsverlag, 2010

[33] Cruse, D. A.: *Lexical Semantics*. Cambridge University Press, 1986

[34] Domínguez García, R. ; Bender, M. ; Anjorin, M. ; Rensing, C. ; Steinmetz, R.: FReSET - An Evaluation Framework for Folksonomy-Based Recommender Systems. In: *Proceedings of the 4th ACM Workshop on Recommender Systems and the Social Web*, 2012

[35] Domínguez García, R. ; P., Scholl ; Rensing, C.: Supporting Resource-based Learning on the Web using Automatically Extracted Large-scale Taxonomies from Multiple Wikipedia Versions. In: *Proceedings of 10th International Conference on Web-Based Learning* Springer, Lecture Notes in Computer Science, Dec 2011. – ISBN 978–3642258121, S. 309–314

[36] Domínguez García, R. ; Rensing, C. ; Steinmetz, R.: Automatic Acquisition of Taxonomies in Different Languages from Multiple Wikipedia Versions. In: *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* ACM International Conference Proceedings Series ACM Inc., ACM International Conference Proceedings Series ACM Inc., Sep 2011. – ISBN 978–1–4503–0732–1

[37] Domínguez García, R. ; Schmidt, S. ; Rensing, C. ; Steinmetz, R.: Automatic Taxonomy Extraction in Different Languages using Wikipedia and minimal language-specific Information. In: *Computational Linguistics and Intelligent Text Processing*, Springer, Mar 2012 (LNCS 7181). – ISBN 978–3642286032, S. 42 – 53

[38] Downes, S.: Learning Objects: Resources for Distance Education Worldwide. In: *The International Review of Research in Open and Distance Learning* 2 (2001), Nr. 1, S. Article–2

[39] Drachsler, H. ; Hummel, H. G. K. ; Koper, R.: Personal Recommender Systems for Learners in Lifelong Learning Networks: the Requirements, Techniques and Model. In: *International Journal of Learning Technology* 3 (2008), Nr. 4, S. 404–423

[40] Dron, J. ; Mitchell, R. ; Siviter, P. ; Boyne, C.: CoFIND – an Experiment in N-dimensional Collaborative Filtering. In: *Journal of Network and Computer Applications* 23 (2000), Nr. 2, S. 131–142

[41] Drummer, J. ; Hambach, S. ; Kienle, A. ; Lucke, U. ; Martens, A. ; Müller, W. ; Rensing, C. ; Schroeder, U. ; Schwill, A. ; Spannagel, C. ; Trahasch, S.: Forschungsherausforderungen des E-Learning. In: *Die 9. E-Learning Fachtagung Informatik* 9 (2011), S. 197–208

[42] Faatz, A.: *Ein Verfahren zur Anreicherung Fachgebietsspezifischer Ontologien durch Begriffsvorschläge*, TU Darmstadt, Diss., Dezember 2004

[43] Felfernig, A. ; Gula, B.: An Empirical Study on Consumer Behavior in the Interaction with Knowledge-Based Recommender Applications. In: *Proceedings of the 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services* IEEE, 2006, S. 37

[44] Felfernig, A. ; Kiener, A.: Knowledge-Based Interactive Selling of Financial Services with FSAdvisor. In: *Proceedings of the National Conference on Artificial Intelligence* Bd. 20 AAAI Press; MIT Press; 1999, 2005, S. 1475

[45] Finin, T. ; Syed, Z.: Creating and Exploiting a Web of Semantic Data. In: *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence*, 2010, S. 7–18

[46] Gabrilovich, E. ; Markovitch, S.: Computing Semantic Relatedness using Wikipedia-Based Explicit Semantic Analysis. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* Bd. 6 Morgan Kaufmann Publishers Inc., 2007, S. 12

[47] Gärtner, K.: *Analyse von Recommendersystemen in Deutschland*. Bd. 38. BIT Verlag, 2012

[48] Goldberg, D. ; Nichols, D. ; Oki, B. M. ; Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry. In: *Communications of the ACM* 35 (1992), Nr. 12, S. 61–70

[49] Granitzer, M. ; Augustin, A. ; Kienreich, W. ; Sabol, V.: Taxonomy Extraction from German Encyclopedic Texts. In: *Proceedings of the Malaysian Joint Conference on Artificial Intelligence*, 2009

[50] Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In: *International Journal of Human-Computer Studies* 43 (1993), Nr. 5, S. 907–928

[51] Hall, M. ; Frank, E. ; Holmes, G. ; Pfahringer, B. ; Reutemann, P. ; Witten, I. H.: The WEKA Data Mining Software: an Update. In: *ACM SIGKDD Explorations Newsletter* 11 (2009), Nr. 1, S. 10–18

[52] Hammwöhner, R.: Qualitätsaspekte der Wikipedia. In: *kommunikation@gesellschaft* 8 (2007), S. 77–90

[53] Hamp, B. ; Feldweg, H.: Germanet – a Lexical-Semantic Net for German. In: *Proceedings of the Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* Citeseer, 1997, S. 9–15

[54] Hearst, M. A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: *Proceedings of the 14th Conference on Computational Linguistics* Bd. 2 Association for Computational Linguistics, 1992, S. 539–545

[55] Herbelot, A. ; Copestake, A.: Acquiring Ontological Relationships from Wikipedia using RMRS. In: *Proceedings of the International Workshop on Web Content Mining with Human Language Technologies* Citeseer, 2006

[56] Herding, D. ; Schroeder, U. ; Stalljohann, P. ; Chatti, M.A.: Formatives Assessment in Offenen, Informellen Vernetzten Lernszenarien. In: *i-com – Zeitschrift für interaktive und kooperative Medien* 11 (2012), Nr. 1, S. 19–21

[57] Herlocker, J. L. ; Konstan, J. A. ; Borchers, A. ; Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. In: *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval* ACM, 1999, S. 230–237

[58] Hodgins, H.W.: The Future of Learning Objects. In: *Proceedings of the 2002 eTEE Conference*, bepress, 2004, S. 11

[59] Hörmann, S.: *Wiederverwendung von digitalen Lernobjekten in einem auf Aggregation basierenden Autorenprozess*, TU Darmstadt, Diss., Februar 2006

[60] Hotho, A. ; Jäschke, R. ; Schmitz, C. ; Stumme, G.: Folkrank: A Ranking Algorithm for Folksonomies. In: *Proceedings of the International Workshop on Information Retrieval (FGIR)* Bd. 2006, Citeseer, 2006

[61] Hotho, A. ; Jäschke, R. ; Schmitz, C.h ; Stumme, G.: BibSonomy: A Social Bookmark and Publication Sharing System. In: *Proceedings of the Conceptual Structures Tool Interoperability Workshop*, 2006

[62] Hsu, M.H.: A Personalized English Learning Recommender System for ESL Students. In: *Expert Systems with Applications* 34 (2008), Nr. 1, S. 683–688

[63] Huang, Y. M. ; Huang, T. C. ; Wang, K. T. ; Hwang, W. Y.: A Markov-Based Recommendation Model for Exploring the Transfer of Learning on the Web. In: *Educational Technology & Society* 12 (2009), Nr. 2, S. 144–162

[64] Jakob, N.: *Extracting Opinion Targets from User-Generated Discourse with an Application to Recommendation Systems*, Technische Universität Darmstadt, Diss., 2011

[65] Jannach, D. ; Zanker, M. ; Felfernig, A. ; Friedrich, G.: *Recommender Systems: An Introduction*. Cambridge University Press, 2010

[66] Janowicz, K. ; Raubal, M. ; Kuhn, W.: The Semantics of Similarity in Geographic Information Retrieval. In: *Journal of Spatial Information Science* 2 (2012), Nr. 2, S. 29–57

[67] Jansen, B. J. ; Spink, A. ; Saracevic, T.: Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. In: *Information Processing & Management* 36 (2000), Nr. 2, S. 207–227

[68] Janssen, J. ; Tattersall, C. ; Waterink, W. ; Berg, B. Van d. ; Van Es, R. ; Bolman, C. ; Koper, R.: Self-organising Navigational Support in Lifelong Learning: How Predecessors Can Lead The Way. In: *Computers & Education* 49 (2007), Nr. 3, S. 781–793

[69] Jäschke, R. ; Marinho, L. ; Hotho, A. ; Schmidt-Thieme, L. ; Stumme, G.: Tag Recommendations in Folksonomies. In: *Knowledge Discovery in Databases: PKDD 2007* 4702 (2007), S. 506–514

[70] Jie, L.: A Personalized E-Learning Material Recommender System. In: *Proceedings of the Second International Conference on Information Technology and Applications*, 2004

[71] Jurafsky, D. ; Martin, J. H. ; Kehler, A. ; Vander Linden, K. ; Ward, N.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Bd. 2. Prentice Hall New Jersey, 2000

[72] Karkada, U.H.: *Friend Recommender System for Social Networks*. Seminararbeit, 2009

[73] Kassner, L. ; Nastase, V. ; Strube, M.: Acquiring a Taxonomy from the German Wikipedia. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008

[74] Khribi, M. K. ; Jemni, M. ; Nasraoui, O.: Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. In: *Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies* IEEE, 2008, S. 241–245

[75] Klein, D. ; Manning, C. D.: Fast Exact Inference with a Factored Model for Natural Language Parsing. In: *Advances in Neural Information Processing Systems* 15 (2002), Nr. 2002, S. 3–10

[76] Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *International Joint Conference on Artificial Intelligence* Bd. 14 Lawrence Erlbaum Associates Ltd, 1995, S. 1137–1145

[77] Koren, Y.: *The Bellkor Solution to the Netflix Grand Prize*. `http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf`. Version: 2009. – Zugriff am 20.11.2012

[78] Koschnick, W. J.: *Standardwörterbuch für die Sozialwissenschaften*. KG Saur Verlag, 1992

[79] Koutrika, G. ; Ikeda, R. ; Bercovitz, B. ; Garcia-Molina, H.: Flexible recommendations over rich data. In: *Proceedings of the 2nd ACM conference on Recommender systems* ACM, 2008, S. 203–210

[80] L'Allier, J. J.: *Frame of Reference: NETg's Map to Its Products, Their Structures and Core Beliefs*. Whitepaper, April 1997

[81] Leacock, C. ; Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: *WordNet: An Electronic Lexical Database* 49 (1998), Nr. 2, S. 265–283

[82] Lehmann, L.: *Lebenszyklusinformationen von Wissensdokumenten - Erfassung, Verwaltung und Validierung*, TU Darmstadt – Multimedia Kommunikation, Diss., April 2010

[83] Lenat, D.B.: CYC: A Large-scale Investment in Knowledge Infrastructure. In: *Communications of the ACM* 38 (1995), Nr. 11, S. 33–38

[84] Littlejohn, A.: *Reusing Online Resources: a Sustainable Approach to E-Learning*. Routledge, 2003

[85] Liu, B.: *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer Verlag, 2007

[86] Lorenzi, F. ; Ricci, F.: Case-Based Recommender Systems: A Unifying View. In: *Intelligent Techniques for Web Personalization* 3169 (2005), S. 89–113

[87] MacKay, D. J. C.: *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ Press, 2003

[88] Maedche, A. ; Staab, S.: Ontology Learning for the Semantic Web. In: *IEEE Intelligent Systems* 16 (2001), Nr. 2, S. 72–79

[89] Maidel, V. ; Shoval, P. ; Shapira, B. ; Taieb-Maimon, M.: Evaluation of an Ontology-Content Based Filtering Method for a Personalized Newspaper. In: *Proceedings of the 2nd ACM conference on Recommender systems* ACM, 2008, S. 91–98

[90] Makhoul, J. ; Kubala, F. ; Schwartz, R. ; Weischedel, R.: Performance Measures for Information Extraction. In: *Proceedings of DARPA Broadcast News Workshop*, 1999, S. 249–252

[91] Manning, C. D. ; Raghavan, P. ; Schutze, H.: *Introduction to Information Retrieval*. Bd. 1. Cambridge University Press Cambridge, 2008

[92] Manouselis, N. ; Costopoulou, C.: Experimental Analysis of Multiattribute Utility Collaborative Filtering on a Synthetic Data Set. In: *Personalization Techniques and Recommender Systems, Series in Machine Perception and Artificial Intelligence* 70 (2008), S. 111–134

[93] Manouselis, N. ; Drachsler, H. ; Vuorikari, R. ; Hummel, H. ; Koper, R.: Recommender Systems in Technology Enhanced Learning. In: *Recommender Systems Handbook*. Springer, 2011, S. 387–415

[94] Martin, P. A.: Semantic Networks to Support Learning. In: *Supplementary Proceedings of International Conference on Computational Science*, Springer, 2008

[95] McCalla, G.: The Ecological Approach to the Design of E-Learning Environments: Purpose-Based Capture and Use of Information about Learners. In: *Journal of Interactive Media in Education* 7 (2004), Nr. 1, S. 18–32

[96] Medelyan, O. ; Milne, D. ; Legg, C. ; Witten, I.H.: Mining Meaning from Wikipedia. In: *International Journal of Human-Computer Studies* 67 (2009), Nr. 9, S. 716–754

[97] Meder, Nobert ; Frick, Andre ; Bettels, Mirko ; Klapschuweit, Christoph: *Web-Didaktik: Eine neue Didaktik webbasierten, vernetzten Lernens*. W. Bertelsmann Verlag, 2002

[98] Melo, G. de ; Weikum, G.: MENTA: Inducing Multilingual Taxonomies from Wikipedia. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* ACM, 2010, S. 1099–1108

[99] Metke-Jimenez, A. ; Raymond, K. ; MacColl, I.: Ontologies Derived from Wikipedia: a Framework for Comparison. In: *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* SciTe Press, 2010, S. 382–387

[100] Meyer, M.: *Modularization and Multi-Granularity Reuse of Learning Resources*, TU Darmstadt, Diss., Oktober 2008

[101] Middleton, S.E. ; Shadbolt, N.R. ; De Roure, D.C.: Ontological User Profiling in Recommender Systems. In: *ACM Transactions on Information Systems* 22 (2004), Nr. 1, S. 54–88

[102] Miller, G.A.: WordNet: a Lexical Database for English. In: *Communications of the ACM* 38 (1995), Nr. 11, S. 39–41

[103] Mitchell, T.: *Machine Learning*. McGraw Hill Higher Education, 1997

[104] Mobasher, B. ; Jin, X. ; Zhou, Y.: Semantically Enhanced Collaborative Filtering on the Web. In: *Web Mining: From Web to Semantic Web* 3209 (2004), S. 57–76

[105] Nakayama, K. ; Hara, T. ; Nishio, S.: Wikipedia Mining for an Association Web Thesaurus Construction. In: *Web Information Systems Engineering* 4831 (2007), S. 322–334

[106] Nastase, V. ; Strube, M. ; Boerschinger, B. ; Zirn, C. ; Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. In: *Proceedings of the International Conference on Language Resources and Evaluation*, European Language Resources Association, 2010. – ISBN 2–9517408–6–7

[107] Naumann, A.: *Wissenserwerb und Informationssuche mit Hypertexten : die Bedeutung von Strukturierung, Navigationshilfen und Arbeitsgedächtnisbelastung*, TU Chemnitz, Diss., 2004

[108] Navigli, R. ; Ponzetto, S. P.: BabelNet: Building a very large multilingual semantic network. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* Association for Computational Linguistics, 2010, S. 216–225

[109] Nguyen, D. ; Matsuo, Y. ; Ishizuka, M.: Subtree Mining for Relation Extraction from Wikipedia. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* Association for Computational Linguistics, 2007, S. 125–128

[110] Oćonnor, M. ; Cosley, D. ; Konstan, J. A. ; Riedl, J.: PolyLens: A Recommender System for Groups of Users. In: *Proceedings of the 2001 Seventh European Conference on Computer Supported Cooperative Work* Springer, 2002, S. 199–218

[111] Page, L. ; Brin, S. ; Motwani, R. ; Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. (1999)

[112] Patwardhan, S. ; Banerjee, S. ; Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2003, S. 241–257

[113] Pazzani, M. ; Billsus, D.: Content-Based Recommendation Systems. In: *The Adaptive Web* 4321 (2007), S. 325–341

[114] Pesquita, C. ; Faria, D. ; Falcão, A. O. ; Lord, P. ; Couto, F. M.: Semantic Similarity in Biomedical Ontologies. In: *PLoS computational biology* 5 (2009), Nr. 7, S. e1000443

[115] Pfeil, U. ; Zaphiris, P. ; Ang, C. S.: Cultural Differences in Collaborative Authoring of Wikipedia. In: *Journal of Computer-Mediated Communication* 12 (2006), Nr. 1, S. 88–113

[116] Polsani, P.R.: Use and Abuse of Reusable Learning Objects. In: *Journal of Digital Information* 3 (2003), Nr. 4, S. 170–179

[117] Ponzetto, S. P. ; Navigli, R.: Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In: *Proceedings of the 21st International Joint Conference on Artifical Intelligence* Morgan Kaufmann, 2009, S. 2083–2088

[118] Ponzetto, S. P. ; Strube, M.: Deriving a Large-Scale Taxonomy from Wikipedia. In: *Proceedings of the National Conference on Artificial Intelligence* Bd. 22 AAAI Press; MIT Press, 2007, S. 1440

[119] Pu, P. ; Chen, L. ; Kumar, P.: Evaluating product search and recommender systems for E-commerce environments. In: *Electronic Commerce Research* 8 (2008), Nr. 1, S. 1–27

[120] Quinlan, J.R.: *C4. 5: Programs for Machine Learning*. Bd. 1. Morgan Kaufmann, 1993

[121] Rafaeli, S. ; Barak, M. ; Dan-Gur, Y. ; Toch, E.: QSIA-a Web-Based Environment for Learning, Assessing and Knowledge Sharing in Communities. In: *Computers & Education* 43 (2004), Nr. 3, S. 273–289

[122] Rafaeli, S. ; Dan-Gur, Y. ; Barak, M.: Social Recommender Systems: Recommendations in Support of E-Learning. In: *International Journal of Distance Education Technologies* 3 (2005), Nr. 2, S. 30–47

[123] Rakes, Glenda C.: Using the Internet as a Tool in a Resource-Based Learning Environment. In: *Educational Technology* 36 (1996), Nr. 5, S. 52–56

[124] Recker, M. ; Walker, A.: Supporting "word-of-mouth" Social Networks through Collaborative Information Filtering. In: *Journal of Interactive Learning Research* 14 (2003), Nr. 1, S. 79

[125] Reichling, T. ; Veith, M. ; Wulf, V.: Expert Recommender: Designing for a Network Organization. In: *Computer Supported Cooperative Work* 16 (2007), Nr. 4, S. 431–465

[126] Rensing, C. ; Bergsträsser, S. ; Hildebrandt, T. ; Meyer, M. ; Zimmermann, B. ; Faatz, A. ; Lehmann, L. ; Steinmetz, R.: Re-Use and Re-Authoring of Learning Resources - Definitions and Examples / TU Darmstadt - Multimedia Communications Lab. Darmstadt, Nov 2005 (KOM-TR-2005-02). – Forschungsbericht

[127] Rensing, C. ; Bogner, C. ; Prescher, T. ; Domínguez García, R. ; Anjorin, M.: Aufgabenprototypen zur Unterstützung der Selbststeuerung im Ressourcenbasierten Lernen. In: *Proceedings of the 9te E-Learning Fachtagung Informatik*. Bonn : Köllen Verlag, Sep 2011. – ISBN 9783885792826, S. 151–162

[128] Rensing, C. ; Böhnstedt, D.: Informelles, Ressourcen-basiertes Lernen. In: *i-com – Zeitschrift für interaktive und kooperative Medien* 11 (2012), May, Nr. 1, S. 15–18. – ISSN 1618–162X

[129] Rensing, C. ; Böhnstedt, D. ; Baumer, C.: Kollaborativer, bedarfsorientierter Wissenserwerb mittels Web-Ressourcen: Prozessmodell, semantische Technologien und eine Communityplattform. In: *Unternehmenswissen als Erfolgsfaktor mobilisieren*. Berlin : Gito GmbH, Sep 2011. – ISBN 9783942183451, S. 241–250

[130] Resnick, P. ; Iacovou, N. ; Suchak, M. ; Bergstrom, P. ; Riedl, J.: GroupLens: an Open Architecture for Collaborative Filtering of Netnews. In: *Proceedings of the ACM conference on Computer Supported Cooperative work* ACM, 1994, S. 175–186

[131] Ricci, F. ; Rokach, L. ; Shapira, B.: Introduction to Recommender Systems Handbook. In: *Recommender Systems Handbook*. Springer, 2011. – ISBN 978–0–387–85820–3, S. 1–35

[132] Richardson, L. ; Ruby, S.: *RESTful Web Services*. O'Reilly Media, Incorporated, 2007

[133] Rijsbergen, C. van: *Information Retrieval*. Cambridge University Press, 1979

[134] Rodenhausen, T.: *Ranking Resources in Folksonomies by Exploiting Semantic and Context-specific Information*, Technische Universität Darmstadt, Diplomarbeit, 2012

[135] Rodenhausen, T. ; Anjorin, M. ; Domínguez García, R. ; Rensing, C. ; Steinmetz, C.: Ranking Resources in Folksonomies by Exploiting Semantic Information. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 2012

[136] Russell, S. J. ; Norvig, P.: *Artificial Intelligence: a Modern Approach*. Prentice hall, 2010

[137] Ryu, P. ; Choi, K.: An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning. In: *Ontology Learning from Text: Methods, Evaluation and Applications* 123 (2005), S. 15

[138] Sanderson, M. ; Croft, B.: Deriving Concept Hierarchies from Text. In: *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval* ACM, 1999, S. 206–213

[139] Santos, O.C.: A Recommender System to Provide Adaptive and Inclusive Standard-Based Support along the eLearning Life Cycle. In: *Proceedings of the 2nd ACM conference on Recommender systems* ACM, 2008, S. 319–322

[140] Sarwar, B. ; Karypis, G. ; Konstan, J. ; Reidl, J.: Item-Based Collaborative Filtering Recommendation Algorithms. In: *Proceedings of the 10th International Conference on World Wide Web* ACM, 2001, S. 285–295

[141] Sarwar, B. M. ; Konstan, J. A. ; Borchers, A. ; Herlocker, J. ; Miller, B. ; Riedl, J.: Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In: *Proceedings of the ACM conference on Computer Supported Cooperative Work* ACM, 1998, S. 345–354

[142] Schein, A. I. ; Popescul, A. ; Ungar, L.H. ; Pennock, D. M.: Methods and Metrics for Cold-Start Recommendations. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ACM, 2002, S. 253–260

[143] Schmidt, S.: *Language-Independent Semantic Relatedness of Web Resources using Wikipedia as ReferenceCorpus*, Technische Universität Darmstadt, Diplomarbeit, 2010

[144] Schmidt, S. ; Scholl, P. ; Rensing, C. ; Steinmetz, R.: Cross-Lingual Recommendations in a Resource-Based Learning Scenario. In: *Proceedings of the 6th European Conference on Technology Enhanced Learning*. Heidelberg : Springer, Sep 2011. – ISBN 9783642239847, S. 356–369

[145] Scholl, P.: *Semantic and Structural Analysis of Web-Based Learning Resources - Supporting Self-directed Resource-based Learning*, TU Darmstadt, Diss., Juni 2011

[146] Scholl, P. ; Böhnstedt, D. ; Domínguez García, R. ; Rensing, C. ; Steinmetz, R.: Extended Explicit Semantic Analysis for Calculating Semantic Relatedness of Web Resources. In: *Proceedings of the 5th European Conference on Technology Enhanced Learning* Bd. Lecture Notes in Computer Science 6383, Springer Verlag, Sep 2010. – ISBN 978–3–642–16019–6, S. 324–339

[147] Schroeder, U. ; Spannagel, C.: Lernen mit Web-2.0-Anwendungen. In: *Navigationen – Zeitschrift für Medien-und Kulturwissenschaften* 8 (2008), Nr. 1, S. 11–18

[148] Seeberg, Cornelia: *Life long learning : modulare Wissensbasen für elektronische Lernumgebungen*. Berlin, TU Darmstadt, Diss., Januar 2003

[149] Shen, L. ; Shen, R.: Learning Content Recommendation Service Based on Simple Sequencing Specification. In: *Proceedings of the International Conference on Web-Based Learning*, Springer, 2004, S. 293–323

[150] Sowa, J. F.: *Principles of Semantic Networks*. Morgan Kaufmann, 1991

[151] Studer, R. ; Benjamins, V. R. ; Fensel, D.: Knowledge Engineering: Principles and Methods. In: *Data & Knowledge Engineering* 25 (1998), Nr. 1, S. 161–197

[152] Suchanek, F. M. ; Kasneci, G. ; Weikum, G.: Yago: a Core of Semantic Knowledge. In: *Proceedings of the 16th International Conference on World Wide Web* ACM, 2007, S. 697–706

[153] Sumida, A. ; Torisawa, K.: Hacking Wikipedia for Hyponymy Relation Acquisition. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing* Citeseer, 2008

[154] Sumida, A. ; Yoshinaga, N. ; Torisawa, K.: Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation Conference*, 2008, S. 28–30

[155] Tang, T. ; McCalla, G.: 'Beyond learners' Interest: Personalized Paper Recommendation Based on their Pedagogical Features for an E-Learning System. In: *PRICAI 2004: Trends in Artificial Intelligence* 3157 (2004), S. 301–310

[156] Tang, T. Y. ; McCalla, G.: Smart Recommendation for an Evolving E-Learning System. In: *Workshop on Technologies for Electronic Documents for Supporting Learning, International Conference on Artificial Intelligence in Education*, 2003, S. 699–710

[157] Tergan, S. O.: Hypertext und Hypermedia: Konzeption, Lernmöglichkeiten, Lernprobleme und Perspektiven. In: *Information und Lernen mit Multimedia und Internet–Lehrbuch für Studium und Praxis* 1 (2002), S. 123–137

[158] Voss, J.: *Collaborative Thesaurus Tagging the Wikipedia way*. 2006. – cs/0604036

[159] Vossen, P.: EuroWordNet: a Multilingual Database for Information Retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, 1997, S. 5–7

[160] Wales, J.: Wikipedia Sociographics. In: *Presentation bei der 21ste CCC-konferenz*, 2004

[161] Walker, A. ; Recker, M. M. ; Lawless, K. ; Wiley, D.: Collaborative Information Filtering: A Review and an Educational Application. In: *International Journal of Artificial Intelligence in Education* 14 (2004), Nr. 1, S. 3–28

[162] Weber, N. ; Buitelaar, P.: Web-Based Ontology Learning with ISOLDE. In: *Proceedings of the Workshop on Web Content Mining with Human Language* Bd. 11 Citeseer, 2006

[163] Wentland, W. ; Knopp, J. ; Silberer, C. ; Hartung, M.: Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. In: *Proceedings of the International Conference on Language Resources and Evaluation*, 2008

[164] Wessner, M.: *Kontextuelle Kooperation in virtuellen Lernumgebungen*. Eul Verlag, 2005

[165] In: Wiley, D.: *Connecting Learning Objects to Instructional Design Theory: a Definition, a Metaphor, and a Taxonomy*. Bloomington, IN. : The Agency for Instructional Technology, 2002, S. 3–23

[166] Willett, P.: The Porter Stemming Algorithm: Then and Now. In: *Program: Electronic Library and Information Systems* 40 (2006), Nr. 3, S. 219–223

[167] Witten, I. H. ; Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005

[168] Witten, I. H. ; Milne, D.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, 2008, S. 25–30

[169] Wong, W.Y.: *Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge*, University of Western Australia, Diss., 2009

[170] Wu, F. ; Hoffmann, R. ; Weld, D. S.: Information Extraction from Wikipedia: Moving Down the Long Tail. In: *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ACM, 2008, S. 731–739

[171] Wu, F. ; Weld, D. S.: Autonomously Semantifying Wikipedia. In: *Proceedings of the 17th ACM Conference on Conference on Information and Knowledge Management* ACM, 2007, S. 41–50

[172] Yamada, I. ; Torisawa, K. ; Kazama, J. ; Kuroda, K. ; Murata, M. ; De Saeger, S. ; Bond, F. ; Sumida, A.: Hypernym Discovery Dased on Distributional Similarity and Hierarchical Structures. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2* Association for Computational Linguistics, 2009, S. 929–937

[173] Zhao, Y. ; Karypis, G. ; Fayyad, U.: Hierarchical Clustering Algorithms for Document Datasets. In: *Data Mining and Knowledge Discovery* 10 (2005), Nr. 2, S. 141–168

[174] Zimmermann, B.: *Pattern-basierte Prozessbeschreibung und -unterstützung: Ein Werkzeug zur Unterstützung von Prozessen zur Anpassung von E-Learning-Materialien*, TU Darmstadt – Elektrotechnik und Informationstechnik – Multimedia Kommunikation, Diss., Dezember 2008

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

ACM	Association for Computing Machinery 5
API	Application Programming Interface 39
CROKODIL	Communities, Web-Ressourcen und Kompetenzentwicklungsdienste integrierende Lernumgebung 43
CSV	Comma-separated values 66
IR	Information Retrieval 8
LOM	Learning Object Metadata 8
MENTA	Multilingual Entity Taxonomy 39
RBL	Ressourcen-basiertes Lernen 1
REST	REpresentational State Transfer Architektur 91
SCORM	Shareable Content Object Reference Model 8
SQL	Standard Query Language 8
URL	Uniform Resource Locator 91
WWW	World Wide Web 1
YAGO	Yet Another Great Ontology 37

ANHANG

A

»Alles nimmt ein gutes Ende für den, der warten kann.«

— Leo Tolstoi

A.1 DETAILS ZU TAXWIKIHEUR.KOM

A.1.1 *Parametrisierung der Heuristiken für die deutsche Sprache*

Für die Parametrisierung der Heuristiken wurde ein anderer Korpus gewählt als in Abschnitt 5.1.4.1, um eine maximale heterogene Linksmenge zu haben. Der Korpus bestand aus 329 zufällig gewählten Kategorien. Der Kategoriengraph selbst bestand aus 635 Links, die manuell gelabelt wurden. Insgesamt bestand der Korpus aus 196 *is-a*-Links und 439 *not-is-a*-Links. Zuerst wird auf die Parametrisierung für das Matching von lexikalischen Köpfen an richtiger Stelle und anschließend auf das Matching von lexikalischen Köpfen an falscher Stelle eingegangen.

A.1.1.1 *Matching von lexikalischen Köpfen an richtiger Stelle*

Um die optimale Konfiguration für die Heuristik zu finden, wurden verschiedene Matching-Fenster in Betracht gezogen. Diese Heuristik wurde auf den Korpus wiederholt angewendet, bis eine optimale Länge des Matching-Fensters bestimmt wurde. Die besten Ergebnisse ergaben sich bei einem Matching-Fenster von 4. 71 von 83 Links wurden dabei korrekt klassifiziert. Die restlichen Ergebnisse sind in folgender Tabelle aufgeführt:

Tabelle 34: Überblick der Ergebnisse bei verändertem Matching-Fenster

Länge des Matching-Fensters	Korrekt kl. Links	Inkorrekt kl. Links
3	73	15
4	71	12
5	59	12
6	46	11

Wie man an den Ergebnissen sieht, wächst die Anzahl der inkorrekt klassifizierten Links, wenn das Matching-Fenster zu klein ist. Für die Matching-Fenster 4, 5 und 6 gibt es einen relativ großen Unterschied in Bezug auf die Anzahl der korrekt klassifizierten Links. Die Anzahl der korrekt klassifizierten Links wächst für kleine Matching-Fenster, aber die Verbesserung von Matching-Fenster 4 zu Matching-Fenster 3 ist nicht signifikant, da zwar 2 Links mehr korrekt klassifiziert wurden, gleichzeitig aber existieren 3 weitere falsch klassifizierte Links.

A.1.1.2 *Matching von lexikalischen Köpfen an falscher Stelle*

Dasselbe Verfahren wurde angewendet, um ein Matching-Fenster für diese Heuristik zu bestimmen. Die besten Ergebnisse wurden mit einem Matching-Fenster von 6 erreicht. 15 von 21 Links wurden korrekt klassifiziert. In Tabelle 35 werden die weiteren Ergebnisse gezeigt.

Tabelle 35: Überblick der Ergebnisse bei verändertem Matching-Fenster

Länge des Matching-Fensters	Korrekt kl. Links	Inkorrekt kl. Links
3	18	9
4	15	8
5	15	8
6	15	6
7	11	5
8	11	5
9	9	5
10	5	1

A.1.2 *Ergebnisse von TaxWikiHeur.KOM in anderen Sprachen*

Tabelle 36: Klassifikationsergebnisse der einzelnen Heuristiken für Englisch

	Ungelabelte Links	Korrekt klassifiziert	Inkorrekt klassifiziert	Noch zu klassifizieren
Heuristik: Admin. Metakategorien	1561	228	0	1333
Heuristik: Verfeinerungslinks	1333	233	0	1100
Heuristik: Lex. Köpfe an richtiger Stelle	1100	94	1	1005
Heuristik: Lex. Köpfe an falscher Stelle	1005	46	0	959
Heuristik: Erster Satz eines Artikels	959	15	4	940
Matching von Kookkurrenzen im Graph	940	75	47	818
Heuristik: Transitive Links		117	89	

Tabelle 37: Klassifikationsergebnisse der einzelnen Heuristiken Spanisch

	Ungelabelte Links	Korrekt klassifiziert	Inkorrekt klassifiziert	Noch zu klassifizieren
Heuristik: Admin. Metakategorien	860	34	0	826
Heuristik: Verfeinerungslinks	826	117	0	709
Heuristik: Lex. Köpfe an richtiger Stelle	709	88	1	620
Heuristik: Lex. Köpfe an falscher Stelle	620	15	4	601
Heuristik: Erster Satz eines Artikels	601	12	0	589
Matching von Kookkurrenzen im Graph	589	32	15	548
Heuristik: Transitive Links		96	28	

Tabelle 38: Klassifikationsergebnisse der einzelnen Heuristiken für Arabisch

	Ungelabelte Links	Korrekt klassifiziert	Inkorrekt klassifiziert	Noch zu klassifizieren
Heuristik: Admin. Metakategorien	1204	256	0	948
Heuristik: Verfeinerungslinks	948	281	0	667
Heuristik: Lex. Köpfe an richtiger Stelle	667	34	5	628
Heuristik: Lex. Köpfe an falscher Stelle	628	72	0	556
Heuristik: Erster Satz eines Artikels	556	51	0	505
Matching von Kookkurrenzen im Graph	505	106	22	377
Heuristik: Transitive Links		28	18	

A.2 DETAILS ZU TAXWIKIML.KOM

A.2.1 *Klassifizierungsergebnisse basierend auf der englischen Wikipedia*

A.2.1.1 *Klassifizierungsergebnisse ohne Einsatz einer Kostenmatrix*

Zur Evaluation wurden 41.177 Links aus dem englischen Wikipedia-Kategoriengraph, die auch in WordNet vorkommen, benutzt. Davon waren 18.977 *is-a*-Links (46,1%) und 22.200 Paare (53,9%) *not-is-a*-Links. Es handelt sich also um ein relativ ausgeglichenes Verhältnis. Ein Überblick der Ergebnisse des Verfahrens wird in Abbildung 39 gezeigt. Ein großer Unterschied zum manuell extrahierten Korpus (siehe Abschnitt 5.2.3.1) war die Tatsache, dass im manuellen Korpus *not-is-a*-Links einen viel größeren Anteil ausgemacht haben. Dennoch war in Englisch, anders als in Deutsch, in beiden Korpora (manuell und automatisch extrahierten) eine größere Anzahl an *not-is-a*-Links beobachtet worden. Das könnte hauptsächlich daran liegen, dass *is-a*-Links eher in WordNet gefunden werden als *not-is-a*-Links. Andererseits ist es auch ein Hinweis darauf, dass die Methode zur Erstellung der Korpora nicht (immer) die Struktur der Wikipedia repräsentativ abbilden kann.

Tabelle 39: Zusammenfassung der Ergebnisse ohne Optimierung

Anzahl der *is-a*-Links im Korpus	18977 (46,09%)
Anzahl der *not-is-a*-Links im Korpus	22200 (53,91%)
Anzahl der korrekt klassifizierten Instanzen	28309 (68,75%)
Anzahl der inkorrekt klassifizierten Instanzen	12868 (21,25%)
Gesamtanzahl der Instanzen	41177

In der evaluierten Menge wurden 28.309 Links (68,75%) korrekt klassifiziert. Die Rate der falsch-negativ klassifizierten Instanzen war relativ niedrig. Im Gegensatz dazu gab es viele falsch-positive Fälle. Hier wurden also viele Paare als *not-is-a* klassifiziert (siehe Tabelle 41). Als Folge davon lag die Genauigkeit für *is-a*-Links bei 60,46% und die Trefferquote für *not-is-a*-Links bei 48%. Man kann also bis zu diesem Punkt sagen, dass die Ergebnisse beim Vergleich mit externen Quellen nicht die erzielte Erkennungsqualität wie beim manuellen Korpus erzielen konnte. Durch die Übergewichtung der Entscheidungen zugunsten von *is-a*-Links konnte eine sehr hohe Trefferquote von 93,02% erreicht werden. Auch die Genauigkeit bei *not-is-a*-Links lag bei hohen 88,95%. Die erreichten F-Maße, 73,2% für *is-a* und 62,3% für *not-is-a*, weisen jedoch auf eine verbesserungsfähige Performance des Klassifikators bei beiden Klassen hin(siehe Tabelle 40).

Bei einem Einsatz im CROKODIL würde der Klassifikator zu einem Konzept eine große Anzahl von verwandten Konzepten zurückliefern, die aber bedingt durch die niedrige Genauigkeit einen hohen Anteil der irrelevanten Konzepte beinhalten, was zu einer Verminderung der Empfehlungsqualität führt. An dieser Stelle wurde der Klassifikator durch eine Kostenmatrix optimiert, so dass falsch-positive Fehler stärker bestraft wurden als andere Fehler. Dieser Prozess wurde im Abschnitt 5.2.3.3 beschrieben und evaluiert.

Tabelle 40: Precision, Recall und F-Maß (Englisch)

Precision	Recall	F-Maß	Klasse
60,5 %	93,0 %	73,3 %	is-a
88,9 %	48,0 %	62,3 %	not-is-a

Tabelle 41: Konfusionsmatrix (Englisch)

a	b	← klassifiziert als
17653	11544	a = is-a
1324	10656	b = not-is-a

A.2.2 *Klassifizierungsergebnisse basierend auf der deutschen Wikipedia*

A.2.2.1 *Klassifizierungsergebnisse ohne Einsatz einer Kostenmatrix*

Bei der deutschen Wikipedia wurden 17.682 Links evaluiert, die sowohl im deutschen Kategoriengraph als auch im GermaNet aufgefunden werden konnten. Dabei befanden sich im Korpus 13.728 Einheiten (77,64%) *is-a*-Links und 3.954 Einheiten (22,36%) *not-is-a*-Links (siehe Tabelle 42). Im Vergleich zur englischen Wikipedia waren die *is-a*-Links im automatisch-erstellten Korpus deutlich mehr. Im Vergleich mit dem manuell extrahiertem Korpus konnte ebenfalls ein großer Unterschied bzgl. der prozentualen Quote zwischen *is-a*- und *not-is-a*-Links festgestellt werden: Während im manuell extrahierten Korpus *is-a*- und *not-is-a*-Links ungefähr in gleichen Anteilen vorkamen, wies der automatisch-extrahierte Korpus eine deutliche Übergewichtung von *is-a*-Links auf. Dennoch konnte der Klassifikator in der deutschen Sprache eine hohe Vertrauenswahrscheinlichkeit erreichen, obwohl der Klassifikator, wie bei der englischen Sprache, auch zu vielen falsch-positiven Fehlern tendierte.

Tabelle 42: Zusammenfassung der Ergebnisse ohne Optimierung

Anzahl der *is-a*-Links im Korpus	13728 (77,64%)
Anzahl der *not-is-a*-Links im Korpus	3954 (22,36%)
Anzahl der korrekt klassifizierten Instanzen	14928 (84,43%)
Anzahl der inkorrekt klassifizierten Instanzen	2754 (15,57%)
Gesamtanzahl der Instanzen	17682

Bei den evaluierten Lemmata-Paaren wurden 14.928 Links (84,4%) korrekt klassifiziert (siehe Tabelle 42). Der Klassifikator produzierte Ergebnisse mit einem sehr hohen Recall von 95,3% bei der Erkennung von *is-a*-Links. Die Precision der Erkennung von *is-a*-Links war mit 86,1% ebenfalls gut. Die Klassifizierung von *not-is-a*-Links ist dagegen deutlich schlechter ausgefallen. Mit 46,6% wurde ein relativ geringer Recall erzielt. Das Ungleichgewicht zwischen der Erkennungsqualität von *is-a* und *not-is-a* kommt im balancierten Maß F-Maß nochmal zum Ausdruck. Während *is-a*-Links mit 90,5% ein vergleichsweise hohes F-Maß aufweisen, fällt mit 57,2% das Ergebnis bei *not-is-a*-Links ungleich niedriger aus. Beim Einsatz in der CROKODIL-Plattform hätte das zur Folge, dass die irrelevanten Ergebnisse niedriger ausgefiltert werden und so in die Trefferliste gelangen könnten. Dennoch soll im Auge behalten werden, dass bei der deutschen Sprache auch ohne den Einsatz von Kostenmatrix eine Vertrauenswahrscheinlichkeit von ca. 85% erreicht werden konnte. Somit konnten die in 5.3.1

aufgeführten Annahmen hinsichtlich der Inhaltsqualität von deutscher Wikipedia aber auch in Hinsicht auf die insgesamt gute Performance von TaxWikiML.KOM in der deutschen Sprache untermauert werden.

Tabelle 43: Precision, Recall und F-Maß (Deutsch)

Precision	Recall	F-Maß	Klasse
86,1 %	95,3 %	90,5 %	is-a
74,1 %	46,6 %	57,2 %	not-is-a

Tabelle 44: Konfusionsmatrix (Deutsch)

a	b	← klassifiziert als
13085	2111	a = is-a
643	1843	b = not-is-a

A.2.2.2 *Klassifizierungsergebnisse mit Einsatz einer Kostenmatrix*

Für die deutsche wurde empirisch folgende Kostenmatrix bestimmt:

$$\begin{pmatrix} 0 & 8 \\ 1 & 0 \end{pmatrix}$$

Genau wie bei der englischen Sprache mussten falsch-positive Fehler stärker bestraft werden. Im Gegensatz zur englischen Sprache (vgl. 5.2.3.3) waren die Ergebnisse von WikiNet geringfügig besser als die Ergebnisse von TaxWikiML.KOM für die überprüften Links. Insgesamt konnte WikiNet 93,44% der Links korrekt klassifizieren (s. Tabelle 45). TaxWikiML.KOM konnte 92,84% der Links korrekt klassifizieren. Darüber hinaus wurde mit der erreichten Vertrauenswahrscheinlichkeit der größte Wert im zwischensprachlichen Vergleich erreicht.

Tabelle 45: Vergleich der Ergebnisse zwischen TaxWikiML.KOM und WikiNet

	TaxWikiML.KOM	WikiNet
Korrekt klass. Links	16416 (92,84%)	16522 (93,44%)
Inkorrekt klass. Links	1266 (7,16%)	1160 (6,56%)
Gesamtanzahl der Links	17682	17682

Tabelle 46 zeigt detaillierte Ergebnisse beider Ansätze für den Evaluationskorpus. Für F-Maß lagen die Werte für *is-a*-Beziehungen bei 95,34 % bei TaxWikiML.KOM und 95,71 % bei WikiNet. Für die deutsche Sprache waren die Klassifikationsergebnisse sehr hoch und gleichzeitig sehr ähnlich. Es lässt sich also keine eindeutige Tendenz erkennen, welcher Ansatz besser abschneidet. Leider konnten nur 17.682 evaluiert werden. Die Gründe wurden schon in 5.2.3.3 besprochen: Wikipedia hat eine viel größere Konzept-Abdeckung als GermaNet und WikiNet und die Tatsache, dass viele Wikipedia-Kategorien nicht einem GermaNet-Synset zugeordnet werden können.

Tabelle 46: Detaillierte Ergebnisse von TaxWikiML.KOM und WikiNet im Vergleich zu GermaNet

	Precision	Recall	F-Maß	Klasse
TaxWikiML.KOM	96,30 %	94,41 %	95,34 %	is-a
	81,82 %	87,41 %	84,52 %	not-is-a
WikiNet	97,18 %	94,29 %	95,71 %	is-a
	82,03 %	90,51 %	86,06 %	not-is-a

A.3 DETAILS ZUR EVALUATION DER NUTZUNG DER TAXONOMIE IM ANWENDUNGSSZENARIO

In diesem Abschnitt werden weitere Einzelheiten zu den im Rahmen der Evaluation dieser Arbeit verwendeten Korpora vorgestellt.

A.3.1 *Weitere Details zu den in der Evaluation verwendeten Korpora*

In diesem Kapitel wird die Dichte der Korpora analysiert. Während im ersten Teil dieses Abschnitts die Dichte der Knoten vor dem Hinzufügen von Hyponymierelationen in der Folksonomie betrachtet wird, werden im zweiten Teil die Daten nach dem Hinzufügen dargestellt. Im jedem Abschnitt werden $d_{min,res}$, $d_{max,res}$, $d_{avg,res}$, $d_{med,res}$ für jeden Knoten-Typ analysiert. Dabei entspricht $d_{min,res}$ / $d_{min,tag}$ / $d_{min,user}$ dem kleinsten Grad einer Ressource, eines Tags oder Benutzers. $d_{max,res}$ / $d_{max,tag}$ / $d_{max,user}$ entspricht dem größten Grad einer Ressource, eines Tags oder Benutzers. Der Durchschnitt der Grade aller Ressourcen, Tags oder Benutzer wird durch $d_{avg,res}$ / $d_{avg,tag}$ / $d_{avg,user}$ dargestellt. Schließlich repräsentiert $d_{med,res}$ / $d_{med,tag}$ / $d_{med,user}$ den Median der Grade aller Ressourcen, Tags oder Benutzer.

A.3.1.1 *Details zu den verschiedenen Knoten-Typen in den Korpora vor dem Hinzufügen der benutzten Korpora*

Tabelle 47: Dichte der Ressource-Knoten in den benutzten Korpora

Datensatz	$d_{min,res}$	$d_{max,res}$	$d_{avg,res}$	$d_{med,res}$
$CROK_p$	1	112	14.17	3
$CROK_1$	1	42	15.05	14
$CROK_2$	1	45	8.28	4

Tabelle 48: Dichte der Tags-Knoten in den benutzten Korpora

Datensatz	$d_{min,tag}$	$d_{max,tag}$	$d_{avg,tag}$	$d_{med,tag}$
$CROK_p$	1	165	9.59	2
$CROK_1$	1	149	18.29	8
$CROK_2$	1	36	5.63	2

Tabelle 49: Dichte der Benutzer-Knoten in den benutzten Korpora

Datensatz	$d_{min,user}$	$d_{max,user}$	$d_{avg,user}$	$d_{med,user}$
$CROK_p$	2	296	130.77	155
$CROK_1$	17	232	70.87	50
$CROK_2$	1	29	15.14	20

A.3.1.2 *Details zu den verschiedenen Knoten-Typen in den Korpora nach dem Hinzufügen von Hyponymierelationen*

Tabelle 50: Dichte der Ressource-Knoten in den benutzten Korpora

Datensatz	$d_{min,res}$	$d_{max,res}$	$d_{avg,res}$	$d_{med,res}$
$CROK_p$	1	112	14.54	4
$CROK_1$	1	52	16.04	14
$CROK_2$	1	45	8.35	4

Tabelle 51: Dichte der Tags-Knoten in den benutzten Korpora

Datensatz	$d_{min,tag}$	$d_{max,tag}$	$d_{avg,tag}$	$d_{med,tag}$
$CROK_p$	1	165	9.84	2
$CROK_1$	1	149	19.49	8
$CROK_2$	1	36	5.67	2

Tabelle 52: Dichte der Benutzer-Knoten in den benutzten Korpora

Datensatz	$d_{min,user}$	$d_{max,user}$	$d_{avg,user}$	$d_{med,user}$
$CROK_p$	2	301	134.14	160
$CROK_1$	17	237	75.54	55
$CROK_2$	1	29	15.27	20

A.3.2 *Weitere Details zu Ausführung von FolkRank auf die verwendeten Korpora*

A.3.2.1 *Ergebnisse von FolkRank mit anderen Parametern ohne Hyponymien*

$CROK_p$:

Tabelle 53: Ergebnisse für $CROK_p$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1285	0,0425	0,0556
2	0,0989	0,0357	0,043
3	0,1168	0,0535	0,0586
4	0,1384	0,1161	0,0898
5	0,1548	0,1615	0,1137
6	0,161	0,1881	0,1266
7	0,1679	0,2124	0,1386
8	0,1688	0,2268	0,1454
9	0,1705	0,2386	0,153
10	0,1699	0,2403	0,1574

Tabelle 54: Gesamtergebnisse

Maß	Wert
Precision	0,1475
Recall	0,1512
F_1-Maß	0,108

Tabelle 55: Ergebnisse für $CROK_p$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1285	0,0425	0,0556
2	0,1017	0,0414	0,0468
3	0,1224	0,0705	0,067
4	0,1412	0,1274	0,0943
5	0,1548	0,1615	0,1137
6	0,161	0,1881	0,1266
7	0,1679	0,2124	0,1386
8	0,1681	0,2264	0,1449
9	0,1711	0,2391	0,1536
10	0,1699	0,2403	0,1574

Tabelle 56: Gesamtergebnisse

Maß	Wert
Precision	0,1486
Recall	0,1546
F_1-Maß	0,1097

Tabelle 57: Ergebnisse für $CROK_p$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1285	0,0425	0,0556
2	0,1017	0,0414	0,0468
3	0,1224	0,0705	0,067
4	0,1412	0,1274	0,0943
5	0,1548	0,1615	0,1137
6	0,161	0,1881	0,1266
7	0,1679	0,2124	0,1386
8	0,1681	0,2264	0,1449
9	0,1711	0,2391	0,1536
10	0,1699	0,2403	0,1574

Tabelle 58: Gesamtergebnisse

Maß	Wert
Precision	0,1486
Recall	0,1546
F_1-Maß	0,1097

$CROK_1$

Tabelle 59: Ergebnisse für $CROK_1$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3234	0,0662	0,1031
2	0,2711	0,1123	0,145
3	0,2604	0,1603	0,1768
4	0,2264	0,1908	0,1793
5	0,199	0,2149	0,1771
6	0,1954	0,2942	0,1957
7	0,1665	0,2113	0,1535
8	0,1686	0,2367	0,1641
9	0,1742	0,2777	0,1789
10	0,174	0,298	0,1839

Tabelle 60: Gesamtergebnisse

Maß	Wert
Precision	0,219
Recall	0,2027
F_1-Maß	0,1654

Tabelle 61: Ergebnisse für $CROK_1$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3234	0,0662	0,1031
2	0,2711	0,1123	0,145
3	0,2604	0,1603	0,1768
4	0,2264	0,1908	0,1793
5	0,199	0,2149	0,1771
6	0,1963	0,2947	0,1963
7	0,1665	0,2113	0,1535
8	0,1686	0,2367	0,1641
9	0,1742	0,2777	0,1789
10	0,1734	0,2978	0,1836

Tabelle 62: Gesamtergebnisse

Maß	Wert
Precision	0,2191
Recall	0,2027
F_1-Maß	0,1654

Tabelle 63: Ergebnisse für $CROK_1$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3234	0,0662	0,1031
2	0,2711	0,1123	0,145
3	0,2604	0,1603	0,1768
4	0,2276	0,1924	0,1807
5	0,199	0,2149	0,1771
6	0,1963	0,2947	0,1963
7	0,1665	0,2113	0,1535
8	0,1686	0,2367	0,1641
9	0,1742	0,2777	0,1789
10	0,1734	0,2978	0,1836

Tabelle 64: Gesamtergebnisse

Maß	Wert
Precision	0,2192
Recall	0,2029
F_1-Maß	0,1656

$CROK_2$

Tabelle 65: Ergebnisse für $CROK_2$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0398	0,0369	0,0379
2	0,0398	0,049	0,0404
3	0,0398	0,0722	0,0478
4	0,0991	0,1943	0,1277
5	0,1609	0,3998	0,224
6	0,1411	0,4191	0,206
7	0,1304	0,4337	0,1955
8	0,1187	0,4363	0,1822
9	0,1133	0,4614	0,1777
10	0,1237	0,5522	0,1979

Tabelle 66: Gesamtergebnisse

Maß	Wert
Precision	0,0991
Recall	0,2941
F_1-Maß	0,1405

Tabelle 67: Ergebnisse für $CROK_2$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0398	0,0369	0,0379
2	0,0398	0,049	0,0404
3	0,0398	0,0722	0,0478
4	0,0991	0,1943	0,1277
5	0,1598	0,3941	0,2221
6	0,1411	0,4191	0,206
7	0,1304	0,4337	0,1955
8	0,1187	0,4363	0,1822
9	0,1133	0,4614	0,1777
10	0,1237	0,5522	0,1979

Tabelle 68: Gesamtergebnisse

Maß	Wert
Precision	0,099
Recall	0,2935
F_1-Maß	0,1403

Tabelle 69: Ergebnisse für $CROK_2$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0398	0,0369	0,0379
2	0,0398	0,049	0,0404
3	0,0398	0,0722	0,0478
4	0,0991	0,1943	0,1277
5	0,1598	0,3941	0,2221
6	0,1411	0,4191	0,206
7	0,1304	0,4337	0,1955
8	0,1187	0,4363	0,1822
9	0,1133	0,4614	0,1777
10	0,1237	0,5522	0,1979

Tabelle 70: Gesamtergebnisse

Maß	Wert
Precision	0,099
Recall	0,2935
F_1-Maß	0,1403

A.3.2.2 *Ergebnisse von FolkRank mit anderen Parametern mit Hyponymien*

$CROK_p$

Tabelle 71: Ergebnisse für $CROK_p$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1404	0,0449	0,0591
2	0,0909	0,0369	0,0437
3	0,1098	0,0568	0,0606
4	0,142	0,1378	0,102
5	0,1534	0,1621	0,1144
6	0,161	0,1897	0,1277
7	0,1656	0,2083	0,137
8	0,1655	0,2214	0,1426
9	0,16	0,225	0,143
10	0,1651	0,2299	0,1517

Tabelle 72: Gesamtergebnisse

Maß	Wert
Precision	0,1453
Recall	0,1509
F_1-Maß	0,108

Tabelle 73: Ergebnisse für $CROK_p$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1404	0,0449	0,0591
2	0,0938	0,0426	0,0475
3	0,1155	0,0739	0,0691
4	0,142	0,1378	0,102
5	0,1545	0,1623	0,1148
6	0,1619	0,1899	0,1281
7	0,1664	0,2086	0,1373
8	0,1662	0,2216	0,1429
9	0,1606	0,2252	0,1433
10	0,1657	0,2302	0,1521

Tabelle 74: Gesamtergebnisse

Maß	Wert
Precision	0,1467
Recall	0,1534
F_1-Maß	0,1095

Tabelle 75: Ergebnisse für $CROK_p$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,1404	0,0449	0,0591
2	0,0938	0,0426	0,0475
3	0,1155	0,0739	0,0691
4	0,142	0,1378	0,102
5	0,1545	0,1623	0,1148
6	0,1619	0,1899	0,1281
7	0,1664	0,2086	0,1373
8	0,1662	0,2216	0,1429
9	0,1606	0,2252	0,1433
10	0,1657	0,2302	0,1521

Tabelle 76: Gesamtergebnisse

Maß	Wert
Precision	0,1467
Recall	0,1534
F_1-Maß	0,1095

$CROK_1$

Tabelle 77: Ergebnisse für $CROK_1$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3333	0,0561	0,0907
2	0,2917	0,1125	0,1462
3	0,2934	0,1656	0,1871
4	0,2591	0,1887	0,1915
5	0,2354	0,2098	0,1922
6	0,2287	0,2785	0,2083
7	0,2143	0,2498	0,1969
8	0,2138	0,2887	0,2093
9	0,2109	0,3112	0,2151
10	0,2031	0,279	0,2033

Tabelle 78: Gesamtergebnisse

Maß	Wert
Precision	0,2501
Recall	0,2109
F_1-Maß	0,1831

Tabelle 79: Ergebnisse für $CROK_1$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3333	0,0561	0,0907
2	0,2917	0,1125	0,1462
3	0,2934	0,1656	0,1871
4	0,2591	0,1887	0,1915
5	0,2354	0,2098	0,1922
6	0,2287	0,2785	0,2083
7	0,2143	0,2498	0,1969
8	0,2138	0,2887	0,2093
9	0,2109	0,3112	0,2151
10	0,2031	0,279	0,2033

Tabelle 80: Gesamtergebnisse

Maß	Wert
Precision	0,2501
Recall	0,2109
F_1-Maß	0,1831

Tabelle 81: Ergebnisse für $CROK_1$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,3333	0,0561	0,0907
2	0,2917	0,1125	0,1462
3	0,2934	0,1656	0,1871
4	0,2591	0,1887	0,1915
5	0,2354	0,2098	0,1922
6	0,2287	0,2785	0,2083
7	0,2143	0,2498	0,1969
8	0,2138	0,2887	0,2093
9	0,2109	0,3112	0,2151
10	0,2031	0,279	0,2033

Tabelle 82: Gesamtergebnisse

Maß	Wert
Precision	0,2501
Recall	0,2109
F_1-Maß	0,1831

$CROK_2$

Tabelle 83: Ergebnisse für $CROK_2$ mit einfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0618	0,0454	0,05
2	0,0562	0,0601	0,0535
3	0,0487	0,0802	0,0562
4	0,1037	0,1983	0,1322
5	0,1602	0,3896	0,2214
6	0,1462	0,4239	0,2117
7	0,1216	0,4031	0,1817
8	0,1124	0,41	0,1719
9	0,1075	0,4332	0,1679
10	0,122	0,5337	0,1941

Tabelle 84: Gesamtergebnisse

Maß	Wert
Precision	0,1031
Recall	0,2877
F_1-Maß	0,1415

Tabelle 85: Ergebnisse für $CROK_2$ mit fünffacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0618	0,0454	0,05
2	0,0562	0,0601	0,0535
3	0,0487	0,0802	0,0562
4	0,1037	0,1983	0,1322
5	0,1591	0,3839	0,2195
6	0,1462	0,4239	0,2117
7	0,1216	0,4031	0,1817
8	0,1124	0,41	0,1719
9	0,1075	0,4332	0,1679
10	0,122	0,5337	0,1941

Tabelle 86: Gesamtergebnisse

Maß	Wert
Precision	0,1029
Recall	0,2871
F_1-Maß	0,1413

Tabelle 87: Ergebnisse für $CROK_2$ mit zehnfacher Benutzergewichtung

k	Precision	Recall	F_1-Maß
1	0,0618	0,0454	0,05
2	0,0562	0,0601	0,0535
3	0,0487	0,0802	0,0562
4	0,1037	0,1983	0,1322
5	0,1591	0,3839	0,2195
6	0,1462	0,4239	0,2117
7	0,1216	0,4031	0,1817
8	0,1124	0,41	0,1719
9	0,1075	0,4332	0,1679
10	0,122	0,5337	0,1941

Tabelle 88: Gesamtergebnisse

Maß	Wert
Precision	0,1029
Recall	0,2871
F_1-Maß	0,1413
MAE	0,0797

WISSENSCHAFTLICHE ARBEITEN DES AUTORS

B

B.1 VERÖFFENTLICHUNGEN ALS ERSTAUTOR

1. Renato Dominguez Garcia, Matthias Bender, Mojisola Anjorin, Christoph Rensing, und Ralf Steinmetz. *FReSET - An Evaluation Framework for Folksonomy-Based Recommender Systems*. In: *Proceedings of the 4th ACM Workshop on Recommender Systems and the Social Web*. 2012.

2. Renato Dominguez Garcia, Sebastian Schmidt, Christoph Rensing, und Ralf Steinmetz. *Automatic Taxonomy Extraction in Different Languages using Wikipedia and minimal language-specific Information*. In: Alexander Gelbukh, Editor, *Computational Linguistics and Intelligent Text Processing*, LNCS 7181, Seiten 42 – 53. Springer, 2012. ISBN 978-3642286032.

3. Renato Dominguez Garcia, Christoph Rensing, und Ralf Steinmetz. *Automatic Acquisition of Taxonomies in Different Languages from Multiple Wikipedia Versions*. In: Michael Granitzer Stefanie Lindstaedt, Editor, *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM International Conference Proceedings Series ACM Inc., ACM International Conference Proceedings Series ACM Inc., 2011. ISBN 978-1-4503-0732-1.

4. Renato Dominguez Garcia, Philipp Scholl, und Christoph Rensing. *Supporting Resource-based Learning on the Web using automatically extracted Large-scale Taxonomies from multiple Wikipedia Versions*. In: Yiwei Cao Rynson Lau Wolfgang Nejdl Howard Leung, Elvira Popescu, Editor, *Proceedings of 10th Intl. Conf. on Web-based Learning (ICWL'11)*, Seiten 309–314. Springer, Lecture Notes in Computer Science, 2011. ISBN 978-3642258121.

5. Renato Dominguez Garcia, Alexandru Berlea, Philipp Scholl, Doreen Böhnstedt, Christoph Rensing, und Ralf Steinmetz. *Improving Topic Exploration in the Blogosphere by Detecting Relevant Segments*. In: Journal of Universal Computer Science, Editor, *Proceedings of the I-Know 2009*, Seiten 177–188. Verlag der Technischen Universität Graz, Graz, Austria, 2009. ISBN 978-3-851-25-060-2.

6. Renato Dominguez Garcia, Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, und Ralf Steinmetz. *Von Tags zu semantischen Netzen - Einsatz im Ressourcenbasierten Lernen*. In: Nicolas Apostopoulos Andreas Schwill, Editor, *Lernen im digitalen Zeitalter - Workshop-Band - Dokumentation der Pre-Conference zur DeLFI 2009*, Seiten 29–36. Logos, Berlin, 2009. ISBN 9783832522735.

7. Renato Dominguez Garcia, Philipp Scholl, Doreen Böhnstedt, Christoph Rensing, und Ralf Steinmetz. *Towards To an Automatic Web Genre Classification*. Technisches Report TR-2008-10, Multimedia Communications Lab, 2008.

8. Mojisola Anjorin, Thomas Rodenhausen, Renato Domínguez García, und Christoph Rensing. *Exploiting Semantic Information for Graph-based Recommendations of Learning Resources*. In: Carlos Kloos Andrew Ravenscroft, Stefanie Lindstaedt und Davinia Hernández-Leo, Editoren, *21st Century Learning for 21st Century Skills. Proceedings of the 7th European Conference on Technology Enhanced Learning, EC-TEL 2012*, Volume 7563, Seiten 9–22. Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-33262-3.

9. Thomas Rodenhausen, Mojisola Anjorin, Renato Dominguez Garcia, und Christoph Rensing. *Context Determines Content - An Approach to Resource Recommendation in Folksonomies*. In: Werner Geyer Andreas Hotho Bamshad Mobasher, Dietmar Jannach, Editor, *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, Seiten 17–24. ACM, New York, NY, USA, 2012. ISBN 978-1-4503-1638-5.

10. Thomas Rodenhausen, Mojisola Anjorin, Renato Domínguez García, Christoph Rensing, und Ralf Steinmetz. *Ranking Resources in Folksonomies by Exploiting Semantic Information*. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, (i-KNOW '12), Graz, Austria*. ACM, New York, NY, USA, 2012. ISBN 978-1-4503-1242-4.

11. Mojisola Anjorin, Renato Dominguez Garcia, und Christoph Rensing. *CROKODIL: a platform supporting the collaborative management of web resources for learning purposes*. In: *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, Seite 361. SIGCSE ACM Special Interest Group on Computer Science Education, ACM, New York, NY, USA, 2011. ISBN 978-1-4503-0697-3. Poster.

12. Mojisola Anjorin, Christoph Rensing, Kerstin Bischoff, Christian Bogner, Lasse Lehmann, Anna Lenka Reger, Nils Faltin, Achim Steinacker, Andy Lüdemann, und Renato Dominguez Garcia. *CROKODIL - a Platform for Collaborative Resource-Based Learning*. In: Raquel M. Crespo Garcia Fridolin Wild Martin Wolpers Carlos Delgado Kloos, Denis Gillet, Editor, *Towards Ubiquitious Learning, Proceedings of the 6th European Conference on Technology Enhanced Learning, EC-TEL 2011*, LNCS 6964, Seiten 29–42. Springer, Heidelberg, 2011. ISBN 9783642239847.

13. Christoph Rensing, Christian Bogner, Thomas Prescher, Renato Dominguez Garcia, und Mojisola Anjorin. *Aufgabenprototypen zur Unterstützung der Selbststeuerung im Ressourcen-basierten Lernen*. In: Steffen Friedrich Holger Rohland, Andrea Kienle, Editor, *DeLFI 2011 - Die 9. e-Learning Fachtagung Informatik*, Seiten 151–162. Köllen Verlag, Bonn, 2011. ISBN 9783885792826.

14. Philipp Scholl, Doreen Böhnstedt, Renato Dominguez Garcia, Christoph Rensing, und Ralf Steinmetz. *Extended Explicit Semantic Analysis for Calculating Semantic Relatedness of Web Resources*. In: Maren Scheffel Stefanie Lindstädt Vania Dimitrova Martin Wolpers, Paul A. Kirschner, Editor, *Sustaining TEL: From Innovation to Learning and Practice Proceedings of EC-TEL 2010*, Volume Lecture Notes in Computer Science 6383, Seiten 324–339. Springer Verlag, 2010. ISBN 978-3-642-16019-6.

15. Philipp Scholl, Doreen Böhnstedt, Renato Dominguez Garcia, Christoph Rensing, und Ralf Steinmetz. *Anwendungen und Nutzen der Automatischen Erkennung von Web-Genres in persönlichen und Community- Wissensnetzen*. In: Nicolas Apostopoulos Andreas Schwill, Editor, *Lernen im digitalen Zeitalter - Workshop-Band - Dokumentation der Pre-Conference zur DeLFI 2009*, Seiten 37–44. Logos, Berlin, 2009. ISBN 9783832522735.

16. Philipp Scholl, Renato Dominguez Garcia, Doreen Böhnstedt, Christoph Rensing, und Ralf Steinmetz. *Towards Language–Independent Web Genre Detection*. In: ACM, Editor, *WWW '09: Proceedings of the 18th international conference on World wide web*, Seiten 1157–1158. ACM, Madrid, Spain, 2009. ISBN 978-1-60558-487-4.

CURRICULUM VITÆ

C

PERSÖNLICHE INFORMATIONEN

Name	Renato Domínguez García
Geburtsdatum	17. September, 1982
Geburtsort	San José, Costa Rica
Nationalität	Deutsch

SCHUL- UND HOCHSCHULAUSBILDUNG

04/2008–heute	Technische Universität Darmstadt (Darmstadt, Deutschland) Promotionskandidat im Fachgebiet Multimedia Kommunikation (KOM) Fachbereich Elektrotechnik und Informationstechnik
10/2000–07/2007	Technische Universität Braunschweig (Braunschweig, Deutschland) Studium der Informatik
09/1999–08/2000	Universität Hannover (Hannover, Deutschland) Studienkolleg
09/2001	Gimnasio Altair de la Sabana (Sincelejo, Kolumbien) Hochschulaufnahmeprüfung ICFEs (Sincelejo, Kolumbien)
Bis 11/1998	Gimnasio Altair de la Sabana (Sincelejo, Kolumbien) Grundschule und Abitur (Sincelejo, Kolumbien)

BERUFSERFAHRUNG

07/2008–heute	Technische Universität Darmstadt (Darmstadt, Deutschland) Wissenschaftlicher Mitarbeiter am Fachgebiet Multimedia Kommunication (KOM)
09/2004–10/2007	Technische Universität Braunschweig (Braunschweig, Deutschland) Verschiedene Tätigkeiten als studentische Hilfskraft an den Instituten für Theoretische Informatik, Informationssysteme und Datenbanken
10/2003–03/2004	GESIS - Gesellschaft für Informationssysteme GmbH (Salzgitter, Deutschland) Industrie-Praktikum im Rahmen des Studiums

AKTIVITÄTEN IN DER LEHRE

2011	Bachelorpraktikum Informatik (Co-Betreuung)
2010, 2011, 2012	Praktikum Multimedia Kommunikation II (Betreuung)
2010, 2011, 2012	Praktikum Multimedia Kommunikation I (Betreuung)
2009, 2010, 2011, 2012	Seminar Multimedia Kommunikation II (Organisation und Betreuung)
2009, 2010, 2011, 2012	Seminar Multimedia Kommunikation I (Organisation und Betreuung)
2009–heute	Betreuer für Bachelor-, Studien-, Diplom- und Masterarbeiten

SONSTIGES UNIVERSITÄRES ENGAGEMENT

2011–heute	Mitglied im Direktorium des Instituts für Datentechnik, Fachbereich 18, TU Darmstadt
2010–heute	Mitglied im Studienausschuss Elektrotechnik und Informationstechnik, Fachbereich 18, TU Darmstadt
2009–heute	Mitglied der Prüfungskomission, Fachbereich 18, TU Darmstadt

BETREUTE STUDENTISCHE ABSCHLUSSARBEITEN

D

1. Markus Migenda. *Supporting Resource-based Learning using semantic graph-based Recommendations*. Masterarbeit, Technische Universität Darmstadt, Voraussichtlicher Abgabetermin: März 2013.

2. Enrique Congolani. *Evaluación de Sistemas Recomendadores de Contenidos Educativos a través de Estudios de Usuarios*. Tesis de Master, Universidad Nacional de la Plata, Abgabetermin steht noch nicht fest.

3. Tabea Born. *Automatische Restrukturierung von Großtaxonomien mit Hilfe von Wiktionary*. Bachelorarbeit, Technische Universität Darmstadt, September 2012.

4. Gennady Kravchenko. *Entwicklung und Evaluation eines sprachunabhängigen Verfahrens zur Erkennung von taxonomischen Beziehungen basierend auf Wikipedia*. Diplomarbeit, Technische Universität Darmstadt, Mai 2012.

5. Thomas Rodenhausen. *Ranking Resources in Folksonomies by Exploiting Semantic and Context-specific Information*. Masterarbeit, Technische Universität Darmstadt, Januar 2012.

6. Jan Matthias Weinert. *Measuring specificity of Wikipedia articles based on lexical, semantic and structural analysis*. Masterarbeit, Technische Universität Darmstadt, Juli 2011.

7. Matthias Bender. *Konzeption und Entwicklung eines Frameworks zur Evaluation semantischer Recommender Systeme*. Bachelorarbeit, Technische Universität Darmstadt, September 2011.

8. Mario Herdt. *Konzeption und Entwicklung eines Taxonomie-basierten Empfehlungssystems für semantische Folksonomien*. Bachelorarbeit, Technische Universität Darmstadt, September 2011.

9. Evgeniy Gilenko. *Automatische Erkennung von hierarchischen Beziehungen zwischen Konzepten mit Hilfe von Wikipedia*. Diplomarbeit, Technische Universität Darmstadt, Juni 2011.

10. Manuel Wick. *Ein Konzept zur Vereinigung von nicht zusammenhängenden Taxonomien mit Hilfe von Wikipedia*. Bachelorarbeit, Technische Universität Darmstadt, Dezember 2010.

ERKLÄRUNG LAUT §9 DER PROMOTIONSORDNUNG

Ich versichere hiermit, dass ich die vorliegende Dissertation allein und nur unter Verwendung der angegebenen Literatur verfasst habe.

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 2012